

TECHNICAL APPLICATION

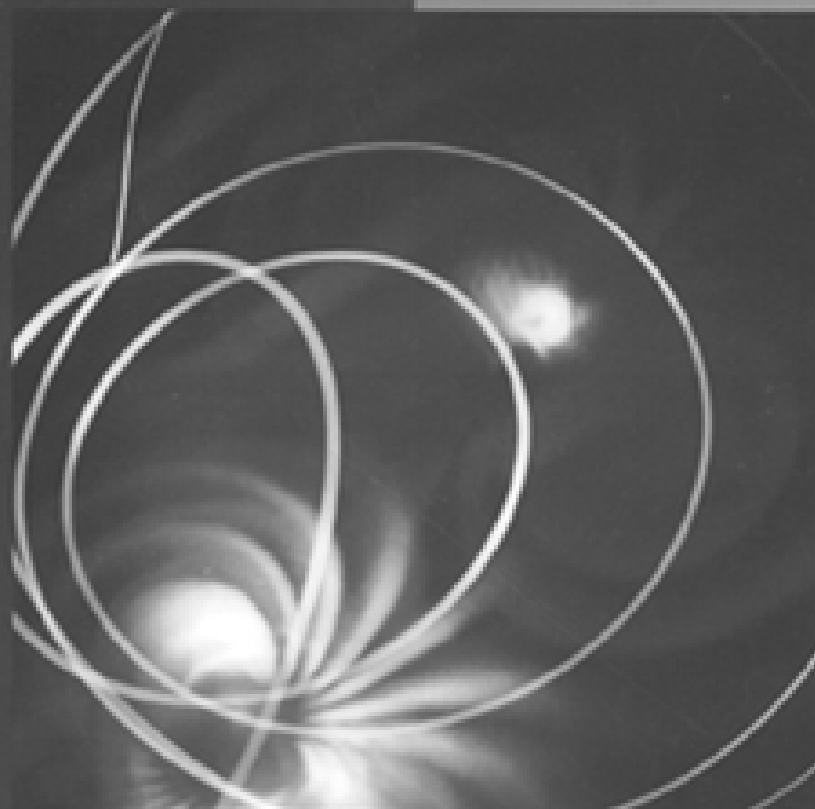
VOLUME 1

NAEP

TWO THOUSAND

2000

National Assessment of Educational Progress



*Application for Cooperative Agreement for the
National Assessment of Educational Progress*

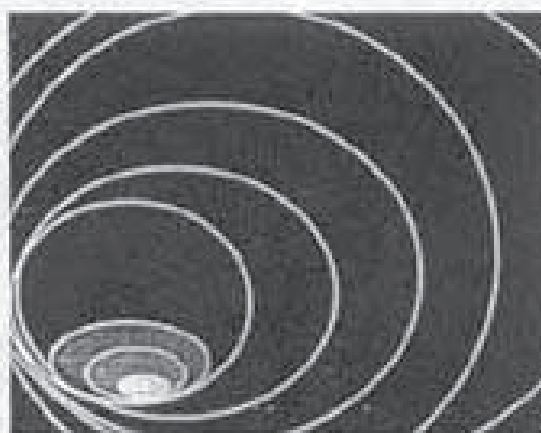
CFDA Number 902 F

July 1, 1998

NAEP

TWO THOUSAND *2000*

National Assessment of Educational Progress



*Application for Cooperative Agreement for the
National Assessment of Educational Progress*

CFDA Number 84.902F

July 1, 1998



Nancy S. Cole
President

Educational Testing Service
Rosdale Road • Princeton, NJ 08541

Telephone: 609-734-1010
Facsimile: 609-921-0235
E-mail: ncs@ets.org
Web Site: www.ets.org

Educational Testing Service



June 17, 1998

Dr. Pascal D. Forgione, Jr.
Commissioner of Education Statistics
National Center for Education Statistics
Education Assessment Division
555 New Jersey Avenue, NW
Washington, DC 20208-5653

Dear Pat:

As President of Educational Testing Service, I am personally excited by the collaborative spirit that has long characterized the NAEP project in general, and which has certainly been evident throughout the development and production of this proposal. The best resources from our research and program management staff have enthusiastically contributed toward creating what we believe is a forward-thinking and high-quality proposal.

Since NAEP became a project at Educational Testing Service in 1983, we have been extremely proud of the role we have played in its history. The program has been marked by commitment to quality and accuracy, and its challenges have been met by the best of our researchers and psychometricians. NAEP is internationally respected, has contributed significantly to American education, and has a reputation that we wish to continue promoting through the ideas presented in this proposal.

ETS brings to the NAEP project its experience, creativity, and deep dedication to American education. We have formed valuable partnerships during the past 15 years of working on NAEP, and the collaborations we have forged for this proposal include some of the best educational and testing organizations in the country. The team of ETS staff and its partners, who will manage the NAEP work, include those who have many years' experience on the NAEP project, as well as new staff who can contribute a fresh perspective to the work. In addition, we will draw on the expertise of countless educators throughout the country with whom we have developed networks of mutual respect and understanding.

NAEP must continue to become more responsive to the varied needs of its constituencies, and this core imperative presents a range of challenges. ETS is committed to continuing to meet these challenges head on, and to offering alternatives that have resulted in efficiencies of both time and cost for the program. Our recent paper

for NCES, *An Integrated Redesign of the National Assessment of Educational Progress*, indicated our willingness and ability both to consider and explore new possibilities for NAEP. In fact, a number of the options we identified have been incorporated into this proposal. Throughout this proposal, efforts to make use of technology and the latest measurement research are evident. We continue, even after 15 years of managing NAEP, to bring new ideas and creative resolutions to the unique challenges of this project.

Our past relationship has been marked by responsiveness to NCES's requests, and we guarantee our continued flexibility in that regard. I would like to highlight some of the specific ideas we present in this proposal, recognizing that NCES's specifications and directives will be met first and foremost. Among the innovations ETS commits to in this proposal are:

- a computer-based testing study,
- a new Web service called "NAEP Interactive," in which members of the public and the press will gain access to customized versions of NAEP data,
- faster reporting schedules,
- parents' report on video, and
- improved Web-based released item packages.

• *Computer-based testing study*

Computer-based testing is at the leading edge of assessment tools, and ETS has already made its mark in the industry. We propose to build on our expertise to explore the use of computer-based testing for NAEP. An analysis of the issues surrounding CBT is a key component of this proposal, and we have outlined some possible options for proceeding with the study.

• *NAEP Interactive*

In conjunction with the production of the summary data tables for the 2000 assessment, ETS is proposing to deliver an innovative product that provides access to NAEP summary results via the World Wide Web. It is called NAEP Interactive, and the product is designed to take advantage of emerging Internet, client/server, and database technologies to rapidly deliver results to users of NAEP data. To fully realize and evaluate the potential of NAEP Interactive, ETS has developed a prototype of the product on CD-ROM and has included it as part of our proposal.

• *Faster reporting schedules*

NAEP reports that are shorter and tailored to specific audiences will be produced. This commitment builds on our current efforts to produce NAEP reports within a shorter timeframe following each assessment.

- *Parents' report on video*

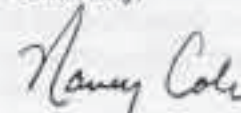
The production of a report for parents is a task required in the proposal. However, ETS proposes to produce such a report on video, using a prominent educator as spokesperson. Such a video could be widely distributed to schools and parent-teacher associations, providing readily available and down-to-earth information about NAEP. Our goal is to make NAEP data more accessible and relevant to the public at large, but in particular to parents.

- *Improved released item packages*

The current released item package is a paper-based product that includes copies of the test items along with item classification information and scoring guides for the constructed-response items. We propose to revise the contents of the package to improve its usability, as well as to make the information available to the public through the World Wide Web.

I will continue to serve as NAEP's cognizant officer, which is the only project overseen by the President at ETS. Because of NAEP's importance to American education and because of its high visibility in light of current national voluntary test efforts, I feel that my time and attention are critical to NAEP. I will marshal all of the resources and talent necessary to ensure NAEP's continued success at ETS. We look forward with excitement and enthusiasm to moving into the 21st century with you.

Sincerely,



Nancy S. Cole
President



June 18, 1998

Dr. Nancy S. Cole, President
Educational Testing Service
Rosedale Road
Princeton, NJ 08541

Dear Dr. Cole:

On behalf of National Computer Systems, Inc., Measurement Services Division, I am pleased to participate with Educational Testing Service as a subcontractor on the Application for Cooperative Agreement for National Assessment of Educational Progress: Development, Scoring, Analysis and Reporting, for 2000 to 2001, CFDA Number 902F.

Our entire organization welcomes the opportunity to continue the successful working relationship established with ETS on the current NAEP activities; it has been an association in which we both may take pride. The accomplishments of this team have meant a very positive contribution to education through the National Assessment of Educational Progress.

We are thoroughly prepared for and fully committed to providing the support services requested by your organization. We look forward to applying our experience with the NAEP team and our expertise in large-scale test processing to deliver the highest quality products and services with maximum effectiveness and efficiency.

Should you need to revisit specific issues, please contact either Brad Thayer or me.

Sincerely,

Gary A. Mainor
Vice President

/nck

Gary A. Mainor

Vice President, Measurement Services Division
National Computer Systems

P.O. Box 30, Iowa City, IA 52244 or 2510 North Dodge Street, Iowa City, IA 52245

Phone: 319-339-6902 Fax: 319-339-6990

E-mail: GMainor@NCS.COM

AMERICAN INSTITUTES FOR RESEARCH



June 17, 1998

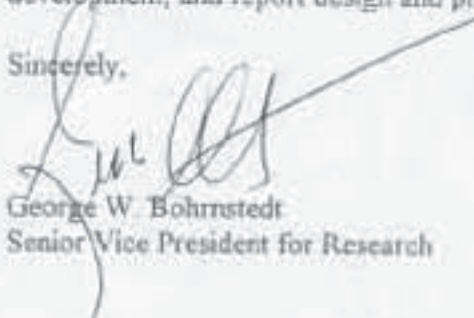
Nancy Cole, President
Educational Testing Service
Rosedale Road
Princeton, NJ 08541-6710

Dear Dr. Cole:

American Institutes for Research (AIR) is pleased to work with ETS as a subcontractor on the proposal Scoring, Analysis, and Reporting of Data: Activities for the National Assessment of Educational Progress (NAEP) from 2000 to 2002.

As you know, we have already forged a successful working relationship in the context of a series of subcontracts to explore school participation incentives and to develop policy and instructional reports from the 1994 and 1996 assessments. We look forward to expanding this relationship under the forthcoming proposal. We also look forward to working closely with you on analysis methodologies, services to states, background questionnaire development, and report design and production.

Sincerely,


George W. Bohrnstedt
Senior Vice President for Research

JOHN C. FLANAGAN RESEARCH CENTER
1791 ARASTRADERO ROAD
P.O. BOX 1113
PALO ALTO, CA 94302
(415) 493-1310



June 11, 1998

Center for the Study of Evaluation
National Center for Research on Evaluation,
Standards, and Student Testing
UCLA Graduate School of Education & Information Studies
201 GSE & IS/Box 951502
Los Angeles, California 90095-1502
(310) 206-1532

Educational Testing Service
Rosedale Road
Mailstop 30-E
Princeton, NJ 08541

Attention: Mr. John Mazzeo
Director of Scoring and Testing

Subject: Proposal for State NAEP Conferences

Dear Mr. Mazzeo:

On behalf of the Regents of the University of California, I am pleased to submit the enclosed proposal for your favorable consideration. The total estimated cost of the award as described in our proposal is \$331,000 for the three year period beginning January 1, 1999, through December 31, 2001. The activities of this project will be conducted under the direction of Professor Eva L. Baker of the UCLA Center for the Study of Evaluation with a subcontract to the University of Colorado at Boulder under the direction of Dr. Robert Linn.

In order to facilitate your review, please contact Dr. Baker at (310) 206-1530 with any questions of a technical nature. If you have any administrative questions, please contact Linda Lee at (310) 825-4032. If an award is made, it should be issued in the name of The Regents of the University of California with the address, Sponsored Research 1401 Ueberroth Building, Mailbox 951406, Los Angeles, CA 90095-1406.

Sincerely,

Eva L. Baker
Director, Center for the Study of Evaluation

Linda Lee
Contract and Grant Officer

ELB/SM:kif

Enclosures: Four originals of the above-referenced proposal

cc w/encl: Rory Constancio, Business Manager
Drusilla Young, Contract and Grant Manager



An Employee-Owned Company

2277 Research Boulevard ♦ Rockville, MD 20850 ♦ (301) 519-5514

Internet: <http://www.aspensys.com>

Eugene A. Brannock
Senior Vice President
General Manager, Information
Services and Technology Group

June 18, 1998

Ms. Nancy Cole
Educational Testing Services
Rosedale Road
Princeton, New Jersey 08541

Reference: Department of Education CFDA No. 902F, National Assessment Of
Educational Progress - Development, Scoring, Analysis and Reporting

Dear Ms. Cole:

We are pleased to participate with Educational Testing Services on the development, scoring, analysis, and reporting of National Assessment of Educational Progress for the 2000, 2001, and 2002 assessments.

We are committed to participating in this project if you are selected as the successful bidder.

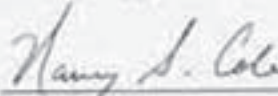
Sincerely,

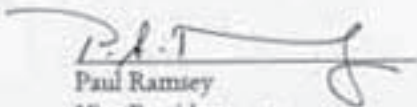
Celebrating Our 40th Anniversary


**APPLICATION FOR COOPERATIVE AGREEMENT
FOR
NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS
2000-2002
CFDA NUMBER 902F**

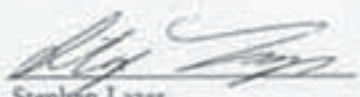
Title: NAEP 2000

Submitted by: Educational Testing Service
Princeton, NJ 08541

Cognizant Officers: 
Nancy S. Cole
President


Paul Ramsey
Vice President

Executive Director: 
Paul L. Williams
Executive Director
(609) 734-1427

Project Director: 
Stephen Lazer
Project Director
(609) 734-1480

Authorized Negotiator: Jacqueline A. Hiller
Assistant Director
Contracts and Grants
(609) 734-5034

Submitted to: Office of Educational Research and Improvement
National Center for Education Statistics
Washington, DC 20208

Period of Work: August 1, 1998 – July 31, 2003

Date Transmitted: July 1, 1998

Status: New Application

NOTICE

The information and data contained herein are proprietary and shall not be duplicated, used, or disclosed other than to evaluate the proposal for possible funding; provided that, if a cooperative agreement is awarded to Educational Testing Service as a result of the submission of this proposal, the Government shall have the right to duplicate, use or disclose the information or data to the extent provided in the cooperative agreement. This restriction does not limit the Government's right to use the information and data obtained from another source without restrictions. The financial information and staff vitae contained in this proposal are considered proprietary data and not required to be released under the Freedom of Information Act.

Any release of information or data under a Freedom of Information Act request will be for the use of the individual requestor only, and will not constitute a license to publish or distribute, or a transfer of any rights in or to the information or data.

Copyright © 1998 Educational Testing Service

Copyright in this material is owned by Educational Testing Service, and limited dissemination has been made for the above purpose only. Except as provided above, no reproduction, use or distribution of any part of this material may be made without the specific authorization of Educational Testing Service.

ACKNOWLEDGMENTS

Primary authors are identified at the beginning of each chapter. However, writing this proposal was largely a collegial effort on the part of the authors and many others who contributed to its preparation. Draft reviews, formal meetings, and informal conversations were ongoing in the evolution of the document into its final form.

had overall responsibility for the proposal development and writing, and he authored the assessment development and special study chapters. managed the proposal process, and wrote the management and staffing sections. coordinated the production effort. prepared the technical assistance reporting chapters, and wrote the operations and scoring chapters. prepared the statistical and analysis chapters, with the support of and . , , and led the writing of the data analysis sections. and led the efforts in the area of World Wide Web applications. produced chapters on advisory committees. wrote the background questionnaire chapters, and contributed to the reporting sections. , working with , , and produced the business application.

Throughout the process, a number of staff contributed through writing and reviews: [REDACTED], [REDACTED], [REDACTED], [REDACTED], [REDACTED], [REDACTED], and [REDACTED]. In addition, [REDACTED] provided overall guidance to the proposal process.

of
 , along with their
respective colleagues, contributed text and information for
significant portions of the proposal, and reviewed numerous
iterations of the text.

oversaw the cover designs and printing
specifications. Expert desktop publishing skills were provided by

Many other staff of ETS and its subcontractors participated in
the conceptualization, writing, reviewing, and production of this
proposal, and we thank them for their efforts. The volumes
comprising this proposal represent the commitment, organizational
support, and team effort that characterize the NAEP project
currently, and which will enable us to carry out the work described
in this Application for Cooperative Agreement.

**APPLICATION FOR COOPERATIVE AGREEMENT
FOR THE
NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS
2000-2002
CFDA NUMBER 84.902 F**

TABLE OF CONTENTS

INTRODUCTION

NAEP REDESIGN AND THIS APPLICATION FOR COOPERATIVE AGREEMENT	I-3
OTHER POSSIBLE COST-SAVING STRATEGIES	I-8
MEETING THE PROGRAM OBJECTIVES	I-9
NOTE ON THE ORGANIZATION OF THIS APPLICATION	I-16

CHAPTER 1. PROJECT INITIATION ACTIVITIES

TASK 1. MEET WITH NCES.....	1-1
TASK 2. TRANSFER AND STORE PROJECT MATERIALS FROM CURRENT RECIPIENT.....	1-2

CHAPTER 2. CONDUCT TECHNICAL REVIEW GROUP MEETINGS

TASK 3. CONDUCT TECHNICAL REVIEW GROUP MEETINGS	2-1
<i>INTRODUCTION</i>	2-1
<i>DESIGN AND ANALYSIS COMMITTEE (DAC)</i>	2-2
<i>TECHNICAL ADVISORY GROUP</i>	2-9
<i>DELIVERABLES</i>	2-12

CHAPTER 3. PREPARE BRIEFING MATERIALS

TASK 4. PREPARE BRIEFING MATERIALS	3-1
<i>BRIEFING MATERIALS</i>	3-1
<i>PRESENT FINDINGS AT ANNUAL MEETINGS</i>	3-2

CHAPTER 4. COORDINATION WITH OTHER NAEP GRANTEES AND CONTRACTORS

TASK 5. COORDINATION WITH OTHER NAEP CONTRACTORS.....	4-1
---	-----

CHAPTER 5. PROVIDE SUPPORT TO NCES, NATIONAL ARCHIVES, AND NAGB AND PROVIDE MONTHLY REPORTS

INTRODUCTION.....	5-1
TASK 6. PROVIDE SUPPORT TO NCES AND THE NATIONAL ARCHIVES	5-2
TASK 7. PROVIDE MONTHLY ADMINISTRATIVE REPORTS	5-4
TASK 8. PROVIDE SUPPORT TO THE NATIONAL ASSESSMENT GOVERNING BOARD.....	5-5

CHAPTER 6. PROVIDE SUPPORT TO STATES

TASK 9. PROVIDE SUPPORT TO STATES.....	6-1
<i>GENERAL TECHNICAL ASSISTANCE</i>	6-2
<i>THE STATE OF EDUCATION</i>	6-3

TABLE OF CONTENTS

<i>NETWORK MEETINGS</i>	6-4
<i>NETWORK NEWS</i>	6-5
<i>ASSISTANCE TO STATES IN THE PROJECTION OF STATE ASSESSMENT SCORES ONTO THE NAEP SCALE</i>	6-6
<i>STATE NAEP CONFERENCES</i>	6-11
<i>OTHER STATE-SUPPORT ACTIVITIES</i>	6-13

CHAPTER 7. DEVELOP AND MAINTAIN STATISTICAL ANALYSIS PROGRAMS

TASK 10. DEVELOP AND MAINTAIN STATISTICAL ANALYSIS PROGRAMS	7-1
<i>OVERVIEW</i>	7-1
<i>PRODUCT DEVELOPMENT AND SUPPORT</i>	7-5
<i>KEY STAFF</i>	7-9

CHAPTER 8. DEVELOP SETS OF BACKGROUND QUESTIONNAIRES

<i>INTRODUCTION</i>	8-1
<i>OVERVIEW OF THE AIR/ETS PLAN FOR THE DEVELOPMENT OF BACKGROUND ITEMS</i>	8-2
<i>CONSTRUCTING RELIABLE, VALID, AND EFFICIENT ITEMS</i>	8-12
<i>SPECIFIC SUBTASKS THAT WILL BE CARRIED OUT AS PART OF TASK 11</i>	8-17

CHAPTER 9. CONDUCT SEMINARS ON PSYCHOMETRICS

TASK 12. CONDUCT SEMINARS ON PSYCHOMETRICS	9-1
<i>INTRODUCTION</i>	9-1
<i>SEMINAR/TRAINING SESSION TOPICS</i>	9-1
<i>SEMINAR/TRAINING SESSION FORMATS</i>	9-3
<i>MEETING ARRANGEMENTS</i>	9-4

CHAPTER 10. FINALIZE THE 2000 NAEP BACKGROUND QUESTIONNAIRES AND PREPARE IMT/OMB CLEARANCE PACKAGE

<i>INTRODUCTION</i>	10-1
TASK 13. FINALIZE THE 2000 NAEP BACKGROUND QUESTIONNAIRES	10-4
<i>CLASSIFY ITEMS INTO SYSTEMATIC FRAMEWORK AND EVALUATE ANALYTICALLY</i>	10-4
<i>USE REPORTING PLANS AND INPUT FROM KEY STAKEHOLDERS TO DRIVE THE FINAL SELECTION OF BACKGROUND ITEMS FOR 2000</i>	10-8
<i>EXPERT TECHNICAL REVIEW BY BACKGROUND COMMITTEE</i>	10-9
TASK 14. PREPARE IMT/OMB CLEARANCE PACKAGE	10-11

CHAPTER 11. PREPARE FOR MATHEMATICS MARKET BASKET

<i>INTRODUCTION</i>	11-1
<i>MARKET BASKETS AND SHORT FORMS: OUR GENERAL APPROACH TO THIS STUDY</i>	11-2
<i>THE ADVANTAGES AND DISADVANTAGES OF MARKET BASKETS AND ISSUES IN NEED OF RESOLUTION</i>	11-5
<i>OVERVIEW OF THE STUDY</i>	11-9
<i>THE DEVELOPMENT OF ITEMS AT ETS</i>	11-13
<i>ETS STANDARDS FOR INSTRUMENT DEVELOPMENT</i>	11-14
<i>THE ETS ASSESSMENT DEVELOPMENT AND REVIEW PROCESS</i>	11-15

TABLE OF CONTENTS

EXERCISE DEVELOPMENT UNDER THIS GRANT	11-26
TASK 15. PREPARE FOR A MARKET-BASKET REPORT IN MATHEMATICS	11-28
STAGE 1: DEVELOP ITEMS FOR STAND-ALONE MARKET-BASKET BOOKLET	11-28
STAGE 2: PILOT TEST SHORT-TEST BOOKLET, ANALYZE PILOT-TEST RESULTS, AND SELECT EXERCISES FOR SPECIAL STUDY	11-41
STAGE 3: ADMINISTER SPECIAL STUDY	11-48
STAGE 4: SCORE AND ANALYZE MARKET BASKET	11-49
 CHAPTER 12. NAEP MATHEMATICS, SCIENCE, READING, AND WRITING	
STANDING COMMITTEES	
OVERVIEW OF TASKS 16 AND 17	12-1
TASK 16. NAEP MATHEMATICS COMMITTEE AND NAEP SCIENCE COMMITTEE	12-3
TASK 17. NAEP READING COMMITTEE AND NAEP WRITING COMMITTEE	12-8
 CHAPTER 13. PRINTING, SPIRALING, BUNDLING, PURCHASING, DISTRIBUTION AND RECEIPT CONTROL, AND RETURN OF MATERIALS FOR SCORING THE 2000 ASSESSMENT	
OVERVIEW	13-1
KEY OPERATIONS STAFF	13-1
TASK 18. PRINT READING ASSESSMENT BOOKLETS FOR 2000	13-3
TASK 19. PRINT MATHEMATICS ASSESSMENT BOOKLETS FOR 2000	13-6
TASK 20. PRINT SCIENCE ASSESSMENT BOOKLETS FOR 2000	13-7
TASK 21. PRINT OTHER ASSESSMENT MATERIALS	13-8
QUESTIONNAIRES AND TRACKING FORMS	13-8
FIELD PUBLICATIONS	13-9
TASK 22. SPIRAL AND BUNDLE MATHEMATICS BOOKLETS	13-10
TASK 23. SPIRAL AND BUNDLE SCIENCE BOOKLETS	13-12
TASK 24. SPIRAL AND BUNDLE READING BOOKLETS	13-13
TASK 25. PURCHASE SCIENCE KITS	13-14
A RADICAL COST-SAVING PROPOSAL	13-15
PURCHASE AND QUALITY CONTROL OF SCIENCE KITS	13-16
TASK 26. DISTRIBUTE ASSESSMENT MATERIALS TO SCHOOLS AND RECEIPT CONTROL	13-18
DISTRIBUTION	13-18
RECEIPT CONTROL	13-20
TASK 27. RETURN ASSESSMENT MATERIALS FOR SCORING	13-23
BOOKLET ACCOUNTABILITY	13-23
 CHAPTER 14. SCORING MATHEMATICS, SCIENCE, AND READING FOR THE 2000 ASSESSMENT	
SCORING THE 2000 ASSESSMENT	14-1
ETS AND NCS CAPABILITIES	14-2
KEY SCORING STAFF	14-3
SCORING TIMELINE	14-5
SCORER QUALIFICATIONS	14-5
IMAGE SCORING	14-7

TABLE OF CONTENTS

SCORING	14-8
TASK 28. SCORE GRADE 4 MATHEMATICS BOOKLETS.....	14-15
TASK 29. SCORE GRADE 8 MATHEMATICS BOOKLETS.....	14-16
TASK 30. SCORE GRADE 12 MATHEMATICS BOOKLETS.....	14-17
TASK 31. SCORE GRADE 4 SCIENCE BOOKLETS.....	14-18
TASK 32. SCORE GRADE 8 SCIENCE BOOKLETS.....	14-19
TASK 33. SCORE GRADE 12 SCIENCE BOOKLETS.....	14-20
TASK 34. SCORE GRADE 4 READING BOOKLETS	14-21

CHAPTER 15. SCAN AND EDIT BACKGROUND QUESTIONNAIRES AND PRODUCE WORKING DATA FILE

OVERVIEW OF NAEP DATA PROCESSING	15-1
TASK 35. SCAN AND EDIT BACKGROUND QUESTIONNAIRES.....	15-9
SYSTEM IMPROVEMENTS	15-15
TASK 40. PRODUCE WORKING DATA FILE.....	15-20
DATA FILE PREPARATION AND DOCUMENTATION	15-20
RESPONDENT-LEVEL DATA FILES.....	15-21
ITEM-LEVEL DATA FILES.....	15-23
MAINTENANCE OF NECESSARY DATA FILES	15-26
COMPUTING RESOURCES	15-30
QUALITY CONTROL.....	15-39

CHAPTER 16. ANALYSIS OF THE NAEP 2000 ASSESSMENTS

INTRODUCTION.....	16-1
SCALING METHODOLOGY FOR THE NAEP 2000 ASSESSMENT DATA.....	16-6
GENERAL CONSIDERATIONS IN THE ESTIMATION OF NAEP POPULATION CHARACTERISTICS.....	16-6
ITEM ANALYSIS (IA) OF COGNITIVE AND BACKGROUND ITEMS.....	16-13
ITEM RESPONSE THEORY (IRT) ESTIMATION OF ITEM PARAMETERS	16-16
LINKING SETS OF ITEMS WITHIN SUBJECT-MATTER AREAS.....	16-22
ESTIMATION OF POPULATION CHARACTERISTICS AND APPROPRIATE STANDARD ERRORS FOR EACH	16-24
LINKING SCALES ACROSS ASSESSMENT YEARS	16-32
PRODUCING SUMMARIES OF THE DATA	16-34
SCALING FOR THE STATE ASSESSMENTS.....	16-36
TECHNICAL DOCUMENTATION OF NAEP SCALING PROCEDURES.....	16-37
METHODOLOGICAL DEVELOPMENTS.....	16-40
TASK 37. SCALING THE NAEP 2000 MATHEMATICS.....	16-45
SCALES AND SAMPLES	16-45
GENERAL CONSIDERATIONS	16-46
ITEM ANALYSES, IRT ESTIMATION OF ITEM PARAMETERS, AND LINKING SETS OF ITEMS WITHIN MATHEMATICS SCALES.....	16-46
ESTIMATION OF POPULATION CHARACTERISTICS AND STANDARD ERRORS, LINKING SCALES ACROSS ASSESSMENT YEARS, AND PRODUCING SUMMARIES OF THE DATA.....	16-47
SCALING FOR THE STATE ASSESSMENTS.....	16-48
ANALYSIS SCHEDULE.....	16-49
DELIVERABLE	16-49

TABLE OF CONTENTS

TASK 38. SCALING THE NAEP 2000 SCIENCE.....	16-50
<i>SCALES AND SAMPLES</i>	16-50
<i>GENERAL CONSIDERATIONS</i>	16-51
<i>ITEM ANALYSES, IRT ESTIMATION OF ITEM PARAMETERS, AND LINKING SETS OF ITEMS WITHIN SCIENCE SCALES</i>	16-51
<i>ESTIMATION OF POPULATION CHARACTERISTICS AND STANDARD ERRORS, LINKING SCALES ACROSS ASSESSMENT YEARS, AND PRODUCING SUMMARIES OF THE DATA</i>	16-53
<i>SCALING FOR THE STATE ASSESSMENTS</i>	16-53
<i>ANALYSIS SCHEDULE</i>	16-54
<i>DELIVERABLE</i>	16-55
TASK 39. SCALING THE NAEP 2000 READING	16-56
<i>SCALES AND SAMPLES</i>	16-56
<i>GENERAL CONSIDERATIONS</i>	16-57
<i>ITEM ANALYSES, IRT ESTIMATION OF ITEM PARAMETERS, AND LINKING SETS OF ITEMS WITHIN READING</i>	16-57
<i>ESTIMATION OF POPULATION CHARACTERISTICS AND STANDARD ERRORS, LINKING SCALES ACROSS ASSESSMENT YEARS, AND PRODUCING SUMMARIES OF THE DATA</i>	16-58
<i>ANALYSIS SCHEDULE</i>	16-58
<i>DELIVERABLE</i>	16-59
TASK 41. CONDUCT DIFFERENTIAL ITEM FUNCTIONING ANALYSIS	16-60
<i>OVERVIEW</i>	16-60
<i>DICHOTOMOUSLY SCORED ITEMS</i>	16-61
<i>POLYTOMOUSLY SCORED ITEMS</i>	16-62
<i>ANALYSIS SCHEDULE</i>	16-62
<i>DELIVERABLES</i>	16-63
TASK 36. ANALYZE TEACHER BACKGROUND DATA	16-64
<i>OVERVIEW</i>	16-64
<i>ALTERNATIVE ONE: SEPARATE ANALYSES OF TEACHER BACKGROUND DATA</i>	16-65
<i>ALTERNATIVE TWO: ANALYSES CONSISTENT WITH PAST NAEP ASSESSMENTS</i>	16-66
<i>ALTERNATIVE THREE: SEPARATE CONDITIONING ON TEACHER BACKGROUND QUESTIONNAIRE DATA</i>	16-67
<i>DEVELOPMENT OF ANALYSIS PLAN AND OUTLINES FOR NATIONAL AND STATE TEACHER REPORT DATA</i>	16-69
<i>ANALYSIS SCHEDULE</i>	16-70
<i>DELIVERABLE</i>	16-70

CHAPTER 17. STORE AND PROTECT COMPLETED INSTRUMENTS AND RECORDS

TASK 42. STORE AND PROTECT COMPLETED INSTRUMENTS AND RECORDS	17-1
<i>MATERIALS STORAGE</i>	17-2

CHAPTER 18. REPORTING OF 2000 RESULTS AND RELEASE OF ITEMS

REPORTING THE RESULTS OF THE NAEP 2000 ASSESSMENT	18-1
<i>GENERAL APPROACH</i>	18-5
TASK 43. REPORT INITIAL RESULTS OF MATHEMATICS ASSESSMENT	18-20
TASK 44. WRITE SHORT, POPULAR REPORT ON MATHEMATICS ASSESSMENT	18-22
TASK 45. REPORT INITIAL RESULTS OF SCIENCE ASSESSMENT.....	18-24
TASK 46. WRITE SHORT, POPULAR REPORT ON SCIENCE ASSESSMENT	18-27
TASK 47. REPORT INITIAL RESULTS OF READING ASSESSMENT	18-29

TABLE OF CONTENTS

TASK 48. WRITE SHORT, POPULAR REPORT ON READING ASSESSMENT.....	18-30
TASK 49. WRITE REPORTS TO INFORM POLICYMAKERS ON 2000 ASSESSMENT.....	18-32
TASK 50. WRITE REPORTS TO INFORM PARENTS ON 2000 ASSESSMENT	18-36
TASK 51. RELEASE ASSESSMENT ITEMS FROM 2000 ASSESSMENT	18-42
TASK 52. ANALYZE AND REPORT ON MATHEMATICS MARKET-BASKET BOOKLET.....	18-45
TASK 53. PROVIDE MATHEMATICS STATE REPORT AND SUPPLEMENTARY	TABLES 18-49
TASK 54. PROVIDE SCIENCE STATE REPORT AND SUPPLEMENTARY TABLES.....	18-53

CHAPTER 19. TRAIN STATE PERSONNEL TO WORK WITH NAEP DATA; PRODUCE SUMMARY DATA TABLES AND DATA PRODUCTS FOR 2000

TASK 55. TRAIN STATE PERSONNEL ON USE OF STATE NAEP DATA	19-1
<i>INTRODUCTION.....</i>	19-1
<i>USING NAEP DATA TO CREATE PRESS RELEASES</i>	19-1
<i>USING NAEP SUMMARY DATA TABLES.....</i>	19-2
<i>USING STATISTICAL SOFTWARE TO ANALYZE NAEP DATA</i>	19-3
<i>ANALYZING TEACHER BACKGROUND DATA</i>	19-4
TASK 56. PRODUCE SUMMARY DATA TABLES FOR MATHEMATICS.....	19-5
<i>SUMMARY DATA TABLES</i>	19-5
<i>REPORTING CONVENTIONS AND SIGNIFICANCE TESTING.....</i>	19-7
<i>DELIVERY OF SUMMARY DATA TABLES</i>	19-8
<i>EXTENSION OF ACHIEVEMENT-LEVEL RESULTS</i>	19-9
<i>NAEP INTERACTIVE: ENHANCED WEB ACCESS TO NAEP RESULTS</i>	19-9
TASK 57. PRODUCE SUMMARY DATA TABLES FOR SCIENCE.....	19-13
TASK 58. PRODUCE SUMMARY DATA TABLES FOR READING.....	19-14
TASKS 60, 61, AND 62. PRODUCE RESTRICTED-USE NATIONAL AND STATE MATHEMATICS AND STATE SCIENCE DATA FILES.....	19-15
<i>DATA FILES AND USER MANUALS</i>	19-15
<i>KEY STAFF.....</i>	19-20
TASK 59. PRODUCE PUBLIC-USE NATIONAL DATA FILE	19-21
<i>KEY STAFF.....</i>	19-22

CHAPTER 20. SUPPORT NAEP'S WORLD WIDE WEB HOME PAGE FOR THE 2000 ASSESSMENT

TASK 63. SUPPORT NAEP'S WORLD WIDE WEB HOME PAGE FOR THE 2000 ASSESSMENT	20-1
<i>INTRODUCTION.....</i>	20-1
<i>THE EVOLUTION OF NAEP'S WORLD WIDE WEB HOME PAGE</i>	20-2
<i>KEY STAFF.....</i>	20-4
<i>DELIVERABLES FOR THE NAEP WEB SITE IN 2000.....</i>	20-5
<i>RE-EVALUATING THE NAEP HOME PAGE</i>	20-11
<i>DESIGNING CONTENT FOR THE WEB—THE POPULAR REPORTS IN 2000</i>	20-13
<i>NAEP TECHNICAL DOCUMENTATION</i>	20-14
<i>PUTTING NEW WEB TECHNOLOGIES TO WORK.....</i>	20-15

TABLE OF CONTENTS

CHAPTER 21. NAEP GEOGRAPHY AND HISTORY STANDING COMMITTEES

TASK 64. NAEP GEOGRAPHY COMMITTEE AND NAEP HISTORY COMMITTEE	21-1
OVERVIEW.....	21-1
PROPOSED NAEP GEOGRAPHY COMMITTEE MEMBERS	21-1
PROPOSED NAEP U.S. HISTORY COMMITTEE MEMBERS	21-3

CHAPTER 22. BACKGROUND QUESTIONNAIRE DEVELOPMENT AND CLEARANCE

INTRODUCTION.....	22-1
TASK 65. REVISE BACKGROUND QUESTIONNAIRE FOR HISTORY AND GEOGRAPHY SUBJECT-SPECIFIC ITEMS.....	22-3
RELATIONSHIP TO THE FRAMEWORK.....	22-3
RELATIONSHIP TO POLICY REPORTS.....	22-4
ITEM PERFORMANCE IN THE 1994 ASSESSMENT.....	22-4
LESSONS FROM THE COGNITIVE LABORATORIES	22-5
PILOT TEST AND ANALYSIS.....	22-5
TASK 66. PREPARE IMT/OMB CLEARANCE PACKAGE FOR 2001 ASSESSMENT.....	22-7
CLEARANCE FOR THE 2000 PILOT TEST.....	22-7
CLEARANCE FOR THE 2001 ASSESSMENT.....	22-8

CHAPTER 23. SPECIAL STUDY OF COMPUTER-BASED TESTING

INTRODUCTION.....	23-1
ISSUES IN USING COMPUTER-BASED TESTING IN NAEP.....	23-3
OVERVIEW OF THE NAEP CBT STUDY	23-18
KEY STAFF.....	23-19
TASK 67. DEVELOP PLAN FOR A SPECIAL STUDY ON COMPUTERIZED TESTING	23-21
STAGE 1: SELECT PROJECT MANAGEMENT TEAM.....	23-21
STAGE 2: MEET WITH NCES TO MAKE FINAL DECISIONS ABOUT STUDY PURPOSE	23-22
STAGE 3: NOMINATE MEMBERS FOR AND CONDUCT MEETINGS OF EXTERNAL CBT DEVELOPMENT COMMITTEE	23-23
STAGE 4: GATHER INFORMATION ON THE AVAILABILITY OF COMPUTERS IN SCHOOLS.....	23-25
STAGE 5: SELECTION OF DELIVERY PLATFORMS	23-25
STAGE 6: SUBMIT PROJECT PLAN	23-27
TASK 68. CONDUCT SPECIAL STUDY IN COMPUTER-BASED TESTING	23-29
STAGE 7: DESIGN AND DEVELOPMENT OF SOFTWARE DELIVERY SYSTEMS	23-29
STAGE 8: SYSTEM TRYOUTS AND PILOTING	23-31
STAGE 9: CONDUCT SPECIAL STUDY OF COMPUTER-BASED TESTING	23-32
STAGES 10 AND 11: ANALYSIS AND REPORTING	23-32

CHAPTER 24. DEVELOPMENT OF REPLACEMENT TASKS FOR U.S. HISTORY AND GEOGRAPHY

INTRODUCTION.....	24-1
THE DEVELOPMENT OF ITEMS AT ETS	24-2
ETS STANDARDS FOR INSTRUMENT DEVELOPMENT	24-4
THE ETS ASSESSMENT DEVELOPMENT AND REVIEW PROCESS.....	24-4
THE ETS/NAEP DEVELOPMENT STAFF	24-16

TABLE OF CONTENTS

TASK 70. DEVELOP REPLACEMENT ITEMS FOR THE 2001 HISTORY	ASSESSMENT	24-18
<i>THE NAEP U.S. HISTORY FRAMEWORK</i>24-18		
<i>DEVELOPMENT OF NEW EXERCISES FOR PILOT TESTING</i>24-22		
<i>THE U. S. HISTORY INSTRUMENT DEVELOPMENT STAFF</i>24-24		
TASK 71. EDIT U.S. HISTORY ITEMS FOR CLARITY		24-26
TASK 72. CONDUCT COGNITIVE LABS ON REPLACEMENT U.S. HISTORY	EXERCISES	24-29
TASK 73. FINALIZE NEW HISTORY ASSESSMENT BLOCKS FOR PILOT TEST		24-31
TASK 74. PREPARE NAGB CLEARANCE PACKAGE FOR HISTORY PILOT TEST		24-32
TASK 75. PREPARE CAMERA-READY COPY FOR HISTORY PILOT TEST	BOOKLETS	24-35
TASK 76. DEVELOP REPLACEMENT ITEMS FOR 2001 GEOGRAPHY	ASSESSMENT	24-38
<i>INTRODUCTION</i>24-38		
<i>THE NAEP GEOGRAPHY FRAMEWORK</i>24-38		
<i>DEVELOPMENT OF NEW EXERCISES FOR PILOT TESTING</i>24-42		
<i>THE GEOGRAPHY INSTRUMENT DEVELOPMENT STAFF</i>24-44		
TASK 77. EDIT NEW GEOGRAPHY ITEMS FOR CLARITY		24-47
TASK 78. CONDUCT COGNITIVE LABORATORIES ON GEOGRAPHY	REPLACEMENT	
ITEMS		24-48
TASK 79. FINALIZE EXERCISES FOR GEOGRAPHY PILOT TEST		24-49
TASK 80. PREPARE NAGB CLEARANCE PACKAGE FOR GEOGRAPHY PILOT	TEST	24-50
TASK 81. PREPARE CAMERA-READY COPY OF GEOGRAPHY PILOT-TEST	BOOKS	24-51

CHAPTER 25. PRINTING AND SHIPPING ASSESSMENT MATERIALS AND RECEIPT CONTROL FOR THE HISTORY AND GEOGRAPHY PILOT TEST

TASK 82. PRINT ASSESSMENT BOOKLETS FOR PILOT TEST		25-1
<i>SPIRAL AND BUNDLE BOOKLETS</i>25-1		
<i>PRODUCE AND PURCHASE ANCILLARY MATERIALS</i>25-2		
TASK 83. SHIP ASSESSMENT MATERIALS FOR PILOT TEST AND RECEIPT	CONTROL	25-3
<i>DISTRIBUTION</i>25-3		
<i>RECEIPT CONTROL</i>25-3		
<i>DATA TAPE CREATION</i>25-4		

CHAPTER 26. SCORING THE PILOT TEST FOR HISTORY AND GEOGRAPHY

TASK 84. SCORING THE PILOT TEST FOR HISTORY AND GEOGRAPHY		26-1
<i>PREPARING FOR SCORER TRAINING</i>26-2		
<i>ON-LINE IMAGE SCORING AND SCORER RELIABILITY MONITORING</i>26-2		

CHAPTER 27. EVALUATION OF PILOT TESTS FOR HISTORY AND GEOGRAPHY

TASK 85. EVALUATE 2001 PILOT-TEST ITEMS		27-1
<i>OVERVIEW</i>27-1		
<i>SAMPLES AND INSTRUMENTS</i>27-1		
<i>ITEM ANALYSES</i>27-2		
<i>DIFFERENTIAL ITEM FUNCTIONING ANALYSES</i>27-5		
<i>DELIVERABLE</i>27-5		

TABLE OF CONTENTS

CHAPTER 28. SELECT, REVISE, AND CLEAR EXERCISES FOR 2001 ASSESSMENTS IN GEOGRAPHY AND U.S. HISTORY

INTRODUCTION.....	28-1
TASK 86. SELECT AND REVISE HISTORY ITEMS	28-2
<i>EVALUATE ITEM NEEDS FOR REPLENISHING U.S. HISTORY BIB.....</i>	<i>28-2</i>
<i>REVIEW PILOT TEST RESULTS.....</i>	<i>28-2</i>
<i>REVIEW ALL EXERCISES AND BLOCKS SELECTED FOR OPERATIONAL USE</i>	<i>28-8</i>
TASK 87. PREPARE NAGB CLEARANCE PACKAGE OF REPLACEMENT HISTORY BLOCKS.....	28-9
TASK 88. PREPARE CAMERA-READY COPY OF REPLACEMENT HISTORY BLOCKS.....	28-12
TASK 89. SELECT AND REVISE GEOGRAPHY ITEMS	28-14
TASK 90. PREPARE NAGB CLEARANCE PACKAGE OF REPLACEMENT GEOGRAPHY BLOCKS.....	28-18
TASK 91. PREPARE CAMERA-READY COPY OF REPLACEMENT GEOGRAPHY BLOCKS.....	28-21

CHAPTER 29. PRINTING, SPIRALING, BUNDLING, PURCHASING, DISTRIBUTION AND RECEIPT CONTROL, AND RETURN OF MATERIALS FOR SCORING FOR THE 2001 ASSESSMENT

TASK 92. PRINT NAEP ASSESSMENT BOOKLETS FOR 2001	29-1
TASK 93. PRINT OTHER ASSESSMENT MATERIALS	29-2
<i>FIELD PUBLICATIONS</i>	<i>29-3</i>
TASKS 94 AND 95. SPIRAL AND BUNDLE U.S. HISTORY AND GEOGRAPHY BOOKLETS.....	29-3
TASK 96. PURCHASE ATLASES AND GEOGRAPHY-RELATED ASSESSMENT MATERIALS.....	29-4
TASK 97. DISTRIBUTE ASSESSMENT MATERIALS TO SCHOOLS AND RECEIPT CONTROL.....	29-5
<i>DISTRIBUTION</i>	<i>29-5</i>
<i>RECEIPT CONTROL</i>	<i>29-5</i>
<i>DATA TAPE CREATION.....</i>	<i>29-5</i>
TASK 98. RETURN ASSESSMENT MATERIALS FOR SCORING	29-6

CHAPTER 30. SCORING U.S. HISTORY AND GEOGRAPHY FOR THE 2001 ASSESSMENT

SCORING THE 2001 ASSESSMENT.....	30-1
<i>PREPARING FOR SCORER TRAINING</i>	<i>30-1</i>
<i>ON-LINE IMAGE SCORING AND SCORER RELIABILITY MONITORING</i>	<i>30-2</i>
TASK 99. SCORE GRADE 4 GEOGRAPHY BOOKLETS	30-4
TASK 100. SCORE GRADE 8 GEOGRAPHY BOOKLETS	30-5
TASK 101. SCORE GRADE 12 GEOGRAPHY BOOKLETS	30-6
TASK 102. SCORE GRADE 4 HISTORY BOOKLETS.....	30-7
TASK 103. SCORE GRADE 8 HISTORY BOOKLETS.....	30-8
TASK 104. SCORE GRADE 12 HISTORY BOOKLETS.....	30-9

*This Application for Cooperative Agreement comprises three volumes.
In addition to this Technical Application, Volume 1, the remaining two
volumes are
Technical Application, Volume 2, and the Business Application.*

INTRODUCTION

The years 2000 through 2002 will be important in the evolution of the National Assessment of Educational Progress (NAEP). This period will be one of transition and innovation, but one in which continuity will remain essential. NAEP will be called on to measure progress in six subject areas, while at the same time laying the groundwork for the transition to the revitalized program envisioned in the National Assessment Governing Board (NAGB) *Policy Statement on Redesigning the National Assessment of Educational Progress*. In these years, a number of important challenges will face the program. Trend lines must be maintained while new, computer-based assessments are introduced. The quality and precision of NAEP reports must remain high, even as they are issued in a more timely fashion. The accuracy of statements in NAEP reports must remain absolute, even as these reports are restructured to more effectively communicate to the general public. While the federal government continues to ensure that NAEP represents a “gold standard” in educational measurement, states must come increasingly to see the program as their own. NAEP must remain the nation’s most respected educational data source, while World Wide Web technologies are used to make its reports, exercises, and data available to vast numbers of people. NAEP must test more subjects more often, while overall program cost remains stable.

While these are formidable challenges, they are welcome. If NAEP is to continue to be the nation’s key educational information system, it must evolve and grow. Systems that were developed to meet the problems and constraints of past decades may no longer be optimal in the new millennium. However, new systems must be introduced in a manner that protects the integrity of the program and its data.

Accomplishing the range of tasks required of NAEP is beyond the scope of any single organization. To meet the varied challenges

of the assessments in 2000, 2001, and 2002, and to provide the appropriate mix of program experience and new ideas, Educational Testing Service (ETS) has assembled a team of unparalleled breadth and expertise. ETS brings 15 years of NAEP instrument development, analysis, and reporting experience, and the staff and processes that have placed NAEP at the forefront of educational assessment programs. National Computer Systems (NCS) will continue to conduct NAEP printing, processing, and scoring. Their robust systems have proven up to the task of managing operational matters for a complex and ever-growing program. American Institutes for Research (AIR) will join the NAEP team under this cooperative agreement. The special expertise of their staff will be used to improve and streamline NAEP background questionnaires, to develop new generations of secondary-user software, to design new cognitive laboratory processes for improving NAEP exercises and questionnaires, to help design reports that will communicate effectively with members of the general public, and to write policy reports based on NAEP data. The Center for Research on Evaluation, Standards, and Student Testing (CRESST) will become a part of the operational NAEP consortium for the first time. They will work with AIR to enable states and others to link assessments to NAEP. CRESST will also help ETS to introduce computer-based testing to the national assessment, and to develop assessments that clearly measure intended constructs. Finally, Aspen Systems will continue to arrange major meetings for NAEP, and will participate in the production of documents both for the field and for the Web.

This team will bring to NAEP a range of psychometric, technical, operational, and assessment development experience never before assembled for any single project. Working together, ETS and its partners will successfully face the range of challenges that will be part of this transitional period.

There is an overall theme that unifies the disparate tasks of this proposal: bringing innovation to NAEP while maintaining the program's trend lines, quality, and integrity. We have selected this theme because we believe that it, more than any other, reflects the challenges and opportunities facing NAEP as it enters a new century.

NAEP Redesign and this Application for Cooperative Agreement

The NAGB *Policy Statement* sets an overall framework for the program's evolution. The National Center for Education Statistics (NCES) has, in its scope of work statement, given operational meaning to that NAGB policy. ETS and its partners have, in this Application for Cooperative Agreement, met NAGB and NCES recommendations wherever possible. The many and varied ways in which we propose to do so are summarized in the paragraphs that follow.

Annual Assessments and Predictable State Assessments

In its *Redesign Policy*, NAGB calls for annual assessments and a predictable and enlarged schedule of state assessments. In this application, concomitant with the stated scope of work, ETS and its partners propose plans to allow for such testing. In 2000, standard assessments in mathematics and science will be conducted nationally at all three grades, and provisions have been made for a reading assessment at grade 4. State assessments at grades 4 and 8 will also occur in mathematics and science. In addition, pilot tests in geography and U.S. history will be conducted, as will a study of a mathematics market-basket form. In 2001, operational national assessments in geography and history will occur at three grades, as well as pilot tests in reading and writing and a special study of computer-based testing. In 2002, national assessments at three grades and state assessments at two grades will be conducted in

reading and writing. This amount of assessment, which has a scope far beyond that ever before implemented in NAEP, will be completed in a manner that allows for analysis and reporting that meets NAGB schedule recommendations. In addition, ETS and its partners have proposed plans that reduce the costs of the science and writing assessments so that such an aggressive testing schedule is possible.

Redesigning Reports to Improve Communication with the General Public

NAGB identifies the general public as the primary audience of NAEP reports. ETS and its partners have proposed an aggressive plan to redesign NAEP reports in support of this goal. ETS and AIR will

how publications should be redesigned. ETS and AIR also propose to work with a leading public opinion firm to design better and more meaningful reports. ETS will produce popular summaries of NAEP data, which can be disseminated to a wide audience.

, ETS proposes to produce a special report based on a market-basket form in mathematics at grade 4.

Making Release of Results More Timely

Consistent with NAGB policy, ETS and its partners propose a plan that will ensure results will be delivered to NCES on a schedule that will allow release of initial assessment results for all subject areas

Linking Assessments to NAEP

NAGB *Redesign Policy* calls for steps to be taken that facilitate the linking of other assessments—especially state testing programs—to NAEP. We propose to meet that goal in several ways. First, staff from AIR and CRESST will provide technical assistance to states interested in linking their assessments to NAEP. Second, we propose to host, under the auspices of CRESST, two special state conferences, one in 2000 and the other in 2002. The first of these conferences will focus on linking issues. In addition, a special study will be conducted to investigate whether NAEP market-basket forms can facilitate linking.

These plans are part of an overall attempt to increase the quality and timeliness of the services we provide to our state partners. We hope, through these efforts, to increase the level of ownership states feel for NAEP, and to support the NCES initiative of having all states participate in NAEP in the year 2000.

Using Technology in NAEP

ETS and its partners propose to make use of technological advances in a number of ways in NAEP. Both the scope of work in the NCES Application for Cooperative Agreement and NAGB policy call for NAEP to begin computer-based testing. Under the guidance of ETS the first operational study of such testing will be conducted on a nationally-representative sample of students in 2001. In addition to testing students, we propose to collect school questionnaire data via the Web in 2001. Such computerized collection should improve both the speed with which we receive data and the quality of those data.

In addition, we propose a number of World Wide Web activities in this proposal that will increase the accessibility and utility of NAEP data and reports. First, ETS proposes to develop a new system called *NAEP Interactive*. This software will allow any

user—even one who knows little about computers or NAEP—to access targeted NAEP data and generate all appropriate tests of statistical significance. Demonstrations of computer-based testing exercises, Web-based surveys, and *NAEP Interactive* can be seen on the compact disk that has been included as a companion to this proposal.

ETS and its partners propose to use the Web in other ways as well. NAEP item release packages will be reconceptualized as Web-based products, in which educators and others can easily access, download, and use items, scoring guides, sample responses, and any appropriate statistical information.

Finally, all NAEP reports will be made available on the Web, and selected publications will be redesigned to take account of the interactive capabilities and non-linear presentation formats available in this medium.

Making NAEP Data More Useful

In the past, the complexities of NAEP have prevented some people from using these data; underlying the NAGB *Redesign Policy* is the notion that NAEP data should be of maximum use to as great a number of people as possible. We propose to make data more usable to a broader range of people in a variety of ways. First, the *NAEP Interactive* system discussed above will make summary data available to even the least sophisticated user. Second, Web and

compact disk versions of the data should ensure that all potential users have easy access. Third, an integrated released-item database will ensure that NAEP exercises play an increasingly important role in American education. Finally, ETS and AIR propose to develop a variety of secondary-user software packages to meet the needs of a range of secondary users. As requested in the NCES scope of work statement, ETS will produce *SPSS*, *SAS*, *S+*, and *Stata* modules that work with data extracted through the *NAEPEX* system. In addition, for users who wish to work with variables that have not been part of the original analysis models, AIR will continue development of its *AM* analysis software. Finally, to facilitate a range of investigations using NAEP data, ETS proposes to continue development of the *NAEPVUE* system previously funded by NCES under a secondary analysis grant. All of these software systems will be made available to the public free of charge.

Rethinking NAEP Background Questionnaires

The NAGB policy calls for background questionnaires that are more focused, yield better data, and reduce respondent burden. Under the leadership of AIR, the plan in this application calls for the use of a modular questionnaire design that involves a short core module to be used in each assessment and targeted modules that will be used only in special situations. The AIR plan also calls for an overhaul of the questionnaire development process,

Streamlining and Improving the Program

ETS and its partners have also proposed operational improvements to NAEP that will reduce costs, increase speed, or improve quality. For example, ETS will work with staff from CRESST and AIR to design new editorial and cognitive laboratory processes that will be

part of NAEP assessment development. These new steps should ensure that fewer items need revision after pilot testing.

Second, ETS and NCS propose a new process for creating camera-ready copy of assessments. This process will reduce by the time needed for photocomposition, and will yield electronic files that support reporting and item release in addition to test booklet production. Related to this, the new ETS Test Creation System will reduce the need for retyping items and will provide electronic records of all exercise-review steps.

Third, as mentioned above, ETS will continue in its efforts to move main NAEP analyses off mainframe computers and onto microcomputers. Use of microcomputers already saves NAEP more than per year.

Fourth, the NCS image archive and retrieval system will eliminate the need for costly rescanning of NAEP booklets for trend scoring. In the future, the system may also obviate the need for warehousing used booklets.

Fifth, NCS will score NAEP constructed-response questions at two sites in 2000 and 2002. This will ensure that the overall scoring effort is finished in half the time it has taken for a comparable amount of work in the past, and will increase the diversity of the scoring pool.

These steps show our commitment to the growth and improvement of NAEP. This commitment is reflected throughout this proposal, as is the willingness of the members of our team to continue, over the life of this cooperative agreement, looking for new efficiencies.

Meeting the Program Objectives

The activities presented in the Application for Cooperative Agreement are program objectives with which ETS and its partners are intimately familiar. We are uniquely qualified to complete these activities with quality, professionalism, enthusiasm, and a true sense of teamwork. Our experienced staffs are recognized as world leaders in assessment development, scoring, and reporting.

The Application for Cooperative Agreement presents six major activities that will be conducted during the project period, which begins in 1998 and ends in 2003. These activities are:

- develop assessment instruments,
- print and assemble assessment booklets,
- score assessment booklets,
- analyze results,

- report assessment results, and
- coordination with other contractors and grantees.

The activities are described in detail below.

Activity 1: Develop Assessment Instruments

For the first time, NAEP pilot tests will be conducted in the same years as operational assessments. This means that this cooperative agreement will have ongoing development activities, which will take a number of forms.

Cognitive subject-area questions for the mathematics market basket will be developed in 1998, pilot tested in 1999, and administered as part of a special study in 2000. The development of the market basket is described in Chapter 11. Questions designed to replace items released after the 1994 NAEP assessments in geography and U.S. history will also be prepared for operational use. As described in Chapters 24 and 28, these will be created in 1999, pilot tested in 2000, and used operationally in 2001. In addition, instruments for a special study of computer-based testing will be developed. Finally, new exercises for the reading and writing assessment will be created in 2000, pilot tested in 2001, and used operationally in 2002 (described in Chapters 38 and 42.)

For each of these efforts, ETS will use an enhanced version of the assessment-development process that has served NAEP well in the past. The proven quality-control steps used in the past will be augmented by cognitive laboratory study and by special editorial reviews to ensure clarity. In addition, the staff members who have developed earlier NAEP instruments will continue to ensure that cognitive questions meet the highest standards of validity, quality, and fairness.

In addition to the cognitive exercises, new background questionnaires will be developed. AIR will play the lead role in this component of the work.

The creation of a new background questionnaire framework, the core background module, and a range of non-core modules, and the development of new analytic models is discussed in Chapters 8, 10, 22, and 37.

Activity 2: Print and Assemble Assessment Booklets

Printing, assembly, and other distribution activities for the 2000, 2001, and 2002 assessments are described in Chapters 13, 29, and 43. Pilot-test printing activities are described in Chapters 11, 25, and 39. ETS and NCS propose to use a streamlined printing process that will support the production of assessment booklets, reports, and Web-based item release packets. Because all the main assessments will be used to measure trend, ETS and NCS will ensure that assessment booklets use the same font, ink color, and order of presentation as was used in past surveys. To ensure that printing, assembly, and distribution work as smoothly and efficiently as possible, ETS and NCS will build on a proven set of production processes and quality-control steps.

Activity 3: Score Assessment Booklets

Scoring the 2000 assessment will represent the largest constructed-response scoring effort in the history of NAEP. In 2000, more than constructed responses will be scored. The 2002 assessment will be of similar size. ETS and NCS are acutely aware of the resources needed to successfully complete a scoring effort of this magnitude.

ETS and NCS will use plans described in Chapters 14, 26, 30, 40, and 44 to ensure that constructed-response scoring of both operational assessments and pilot tests meets the highest standards of quality. Careful recruitment and training plans will be followed. As in recent NAEP assessments, tools available through the NCS image-scoring system will provide real-time data and fine control over the scoring process. Specifically, the image system provides up-to-the-minute data on interrater reliability and trend reliability and validity, and gives managers tools to identify and quickly correct problems.

ETS and NCS are committed to the goal of publishing NAEP reports earlier than ever before, and doing so will necessitate a decrease in the time required for scoring. ETS and NCS have assembled the best combination of staff and physical resources to guarantee rapid, accurate scoring, and the timely delivery of data for analysis and reporting. In addition, NCS will open a second scoring center for NAEP in 2000 and 2002 to ensure that the entire operational scoring occurs within a six-week period (as opposed to 10 weeks in the past), and that pilot-test scoring can be completed within the same time frame.

Activity 4: Analyze Results

ETS will enhance the successful and fully operational approaches used in past assessments to manage the massive amount of data required for the 2000, 2001, and 2002 analyses under the accelerated schedules called for in the Application for Cooperative Agreement. These data systems provide for full documentation of the data files in easy-to-use formats fully responsive to all NAEP data management, analysis, and reporting tasks. The data systems provide for linking across the various components of the booklets, questionnaires, and special studies, as well as for easy linking of the data to previous assessments. Stringent security policies and procedures will be followed. ETS has developed software systems

that provide for all NAEP data conversion and editing requirements. All of the software for the state-of-the-art psychometric procedures proposed for the analyses is already operational. Data processing and management are discussed in Chapters 15, 31, and 45.

ETS introduced scaling into NAEP during the 1984 assessment and has worked diligently during the past decade to refine the procedures to allow for multiple scales, item response theory (IRT) scaling of polytomous student responses, and greater accuracy in secondary analyses. In producing the psychometric scales required for NAEP's 2000, 2001, and 2002 reports, ETS will:

- check the key characteristics of item functioning, through item analysis, DIF analysis, and IRT item parameters;
- use three distinct scaling models to create the various scales specified in the NAGB frameworks for each curriculum area, including the generalized partial-credit model for polytomous items developed by ETS for use with NAEP data;
- apply plausible values technology to permit accurate estimation of population characteristics based on few responses from each student in each scaling area, also developed by ETS for use with NAEP data;
- estimate percentages of students at or above the achievement levels established by NAGB;
- equate results across assessment years in a technically defensible manner to provide short-term and long-term trend results; and
- relate proficiency in each of the assessed curriculum areas to the background questionnaire data obtained from students, teachers, and schools using sophisticated methodological techniques such as hierarchical and multivariate analyses.

The analysis of main assessment results is discussed in Chapters 16, 32, and 46.

In addition to analyzing assessment results, ETS will make final data available in several formats. Restricted-use data files will be created for dissemination on CD-ROM. Public-use data files will be placed both on CD-ROM and on the Internet. Public- and secondary-use data files will be accompanied by a user guide. Tabular summaries will also be available both via CD and the Web; these will be accompanied by an interactive browser (discussed in Chapters 7, 19, 35, and 48, and shown on the CD included with this proposal).

Pilot-test analyses will also be conducted. These will include item analysis for both multiple-choice and constructed-response questions and DIF analysis. Scaling will be included if requested by NCES. Pilot test analyses are described in Chapters 11, 27, and 41.

We also plan to provide full technical documentation of all analyses. The format of this documentation will be the subject of discussion, but we plan to use the Internet to make technical information more accessible and usable.

Activity 5: Report Assessment Results

The purpose of NAEP is to report on educational achievement; all other program activities exist in service of this goal. The improvement and reconceptualization of reporting is at the center of this cooperative agreement. Under NCES and NAGB leadership, ETS has already initiated important changes in NAEP reports. The process that lead to the improvements in the past, characterized by greater planning and broader consumer participation, will be refined and extended for the 2000, 2001, and 2002 reports. ETS and AIR are committed to improving the report development process

and providing reports that are timely, accurate, and useful to the general public and national and state policymakers.

Six goals will characterize the production of the NAEP reports. First, initial results will be prepared to allow for release within of the end of data collection. Second, the formats and contents of reports will be refined in creative ways that will enhance their informative value. Third, special efforts will be made to design reports that communicate effectively with NAEP's primary audiences, members of the general public and national and state policymakers. Fourth, ETS and AIR are committed to "thinking outside the box," and moving beyond traditional print delivery in our efforts to communicate . Fifth, we will use Internet technology in ways that increase the accessibility and utility of NAEP reports. Sixth, we will view exercises released from NAEP as part of our integrated reporting strategy, rather than as separate products for use only by testing experts. Web technology opens broad ranges of possibilities in this area.

Consistent with these goals, we propose a strategy that introduces new and innovative elements into NAEP reporting. The publications we propose, which will be made available both in print and on the World Wide Web, include report cards, state reports, popular summaries, policy reports, and a report on the market-basket study. In addition, consistent with the scope of work statement in the Application for Cooperative Agreement, we propose

We also think that the initial study of computer-based testing in NAEP should be reported on CD-ROM, and that an operational version of the computerized testing instrument should be part of this report.

To ensure that these reports will serve the purposes for which they are designed, we propose extensive planning activities.

In addition, we propose to work with a major public relations firm during this phase. Activities related to report production and item release are described in Chapters 18, 34, and 48.

As mentioned above, the Internet will be a major part of our reporting strategy. All reports and item release kits will be made available via the Web, as will interactive tabular summaries, public-use data products, and technical documentation. Our aggressive Internet strategy is described in Chapters 20, 36, and 49.

All of these approaches, and others that will be planned and discussed with NCES and others, are targeted at the objective of providing the very best and most relevant information in the shortest time possible.

Activity 6: Coordination with Other Contractors and Grantees

As NAEP has become more visible and important, the number of people working with NAEP data has increased exponentially. ETS and its partners are committed to working with such groups in any ways NCES authorizes. Because we have access to key staff at ETS, AIR, and CRESST, we can bring a range of talents to bear in ways that will greatly facilitate the work of users of NAEP data.

We have proposed an aggressive plan for working with other contractors and grantees. As described in Chapter 4, we have made extensive provisions for working with representatives of organizations such as the National Academy of Sciences, the NAEP Validity Studies Panel, and the National Education Goals Panel, as well as for supporting the recipient of the NAEP data collection cooperative agreement. Chapter 5 discusses our plans for supporting NAGB and its contractors, working with the National Archives, and conducting special technical or operational analyses

requested by NCES. In Chapter 6, we outline our plans for providing special support to state participants in NAEP.

Note on the Organization of this Application

This application will contain three volumes: two volumes comprise the technical application, and the third is the business application.

The first and second volumes of the technical application include the ETS plan for conducting work related to the 2000, 2001, and 2002 assessments. The NCES scope of work statement in the Application for Cooperative Agreement included 175 specific suggested tasks. We have slightly redefined the tasks, so that there are now 182. These have been organized into 50 chapters. The first nine chapters deal with introductory and general matters. Chapters 10 through 20 describe work related to the 2000 assessment. 2001 assessment activities are discussed in Chapters 21 through 36. Finally, Chapters 37 through 50 describe work for the 2002 assessment.

CHAPTER 1.

PROJECT INITIATION ACTIVITIES

Tasks 1 and 2
Stephen Lazer

EXECUTIVE SUMMARY

Objective: To ensure that project work begins in an efficient and effective fashion. To meet this objective, Educational Testing Service (ETS) and its partners will:

- convene a project initiation meeting in Washington, DC within two weeks of the award of the cooperative agreement, and
- store, in a secure facility, all materials from previous assessments.

TASK 1.

MEET WITH NCES

To ensure that the project begins in a well-planned and coordinated manner, Educational Testing Service (ETS) will convene an initial meeting with the National Center for Education Statistics (NCES) to develop a mutually acceptable plan for the entire 60 months of the project.

This important full-day meeting, which will occur no later than two weeks after the initiation of the project (by about August 15, 1998), will enable the staff to review and refine the project plans. As the meeting plans are formulated, ETS will discuss with NCES appropriate attendees based on the mutual needs of ETS, NCES, National Computer Systems (NCS), American Institutes for Research (AIR), the Center for Research on Educational Standards and Student Testing (CRESST), and Aspen Systems. We also propose that representatives from the organization holding the data collection cooperative agreement attend this initial meeting. We have made provisions for five ETS staff members and one representative from each of the subcontracting organizations to attend the meeting.

ETS will provide electronic copy of an agenda at least five working days before the scheduled meeting. The agenda will make provision for discussion of the following topics:

- the managerial and organizational structure of the project,
- the scope and schedule of project work, and
- major issues as seen by NCES or ETS and its partners.

ETS will, of course, add to the agenda any other topics requested by NCES.

TASK 2.

TRANSFER AND STORE PROJECT MATERIALS FROM CURRENT RECIPIENT

Since ETS is currently the holder of the cooperative agreements for the development, scoring, analysis, and reporting of the 1996, 1997, and 1998 assessments, no materials must be transferred. NCS will, under this task, store all pre-2000 materials for as long as is indicated by NCES. For this reason, there is a budget associated with this task.

CHAPTER 2.

CONDUCT TECHNICAL REVIEW GROUP MEETINGS

Task 3
James Carlson

EXECUTIVE SUMMARY

Objective: To ensure that the highest caliber of technical expertise is devoted to NAEP design, analysis, and reporting. To achieve this objective, Educational Testing Service (ETS) will:

- continue to manage the work of the NAEP Design and Analysis Committee (DAC),
- conduct formal DAC meetings three times each year,
- consult, on an as-needed basis, with members of the committee via phone, fax, or e-mail for input on specific design and analysis issues,
- maintain the high caliber of the DAC membership when filling vacancies that arise, making a concerted effort to recruit women and members of minority groups as DAC members, and
- create an internal Technical Advisory Group (TAG) consisting of senior-level psychometricians, statisticians, and policy analysts from ETS, American Institutes for Research (AIR), the Center for Research on Evaluation, Standards, and Student Testing (CRESST), and the organization charged with NAEP sampling and data collection.

TASK 3.

CONDUCT TECHNICAL REVIEW GROUP MEETINGS

Introduction

The NAEP Design and Analysis Committee (DAC) has played a major role in advising Educational Testing Service (ETS) and the National Center for Education Statistics (NCES) on issues relating to the design of NAEP and the methodology used in the analysis of NAEP data. As the methodology has evolved under an overarching emphasis on more efficiency, the committee has continually provided recommendations that have helped NAEP maintain state-of-the-art procedures for generating results. We see this role as continuing to be of paramount importance as NAEP evolves.

For NAEP 2000, we propose to add a new additional advisory committee that will serve a role complementary to that of DAC. In recent assessments, there have been several instances of technical issues that arose as ETS staff reviewed results of initial analyses of NAEP data. These issues typically required immediate attention in order that further data analysis could proceed and reports be produced on schedule. To ensure rapid attention to such issues, a Technical Advisory Group (TAG) will be formed, composed of senior psychometricians and statisticians from the staff of ETS, American Institutes for Research (AIR), and the Center for Research on Evaluation, Standards and Student Testing (CRESST). The members of TAG will be familiar with NAEP methodology but not directly involved in the day-to-day operational analyses. They will be available for consultation via e-mail and teleconferencing whenever a problem arises.

Design and Analysis Committee (DAC)

For as long as ETS has participated in the NAEP project, the DAC has played an integral role in the work. The DAC has helped ETS staff develop world-class methodology that is recognized nationally and internationally as a standard for assessment excellence. This outstanding group of statistical, measurement, sampling, and policy analysis experts has been assembled to provide timely advice on important NAEP issues. DAC members serving under the current agreement will continue through the year 2000. It is anticipated that most of them, listed below, will continue to serve should ETS be awarded the cooperative agreement for the assessments covered under this application.

The DAC has played a leading role in the development of NAEP design, analysis, and reporting plans. Recommendations from this committee have, in particular, led to the adoption of Item Response Theory (IRT) and plausible-values methodologies developed by ETS psychometricians. The DAC was instrumental in designing the study conducted by ETS research scientists that verified that the plausible-values technique produced consistent results. Also, members of DAC took a proactive role in the development of enhanced methodology in several areas, such as differential item functioning analyses (DIF) and methods for performing multiple comparisons that are used in reporting NAEP results.

The DAC will continue to play an active role in NAEP design and analysis activities, including but not limited to:

- reviewing and commenting on overall project plans,
- identifying research questions,
- consulting on design of special studies related to NAEP methodology,

CHAPTER 2. CONDUCT TECHNICAL REVIEW GROUP MEETINGS

- providing input about instrument development,
- proposing analytical models and methods,
- reviewing work plans and their implementation, and
- reviewing reports, especially those of a technical nature.

The DAC will meet _____ over the life of this cooperative agreement (August 1, 1998 through July 31, 2003). However, DAC meetings until July 1, 2000 will be administered by ETS under the pre-existing cooperative agreement. Thus, this Application contains budgetary provision for DAC activities between July 1, 2000 and July 31, 2003. Typically, each meeting will commence on a Thursday evening, include a full-day Friday session, and conclude in the middle of the day on Saturday. All meetings will be held in Washington, DC, unless the NCES project officer grants prior approval for meeting at a different location.

ETS will arrange for meeting space and lodging for participants in each DAC meeting. It is anticipated that participants will include four to six staff members from NCES, one to four from the National Assessment Governing Board (NAGB), and four to six from other NAEP contractors and subcontractors.

ETS staff will prepare a preliminary agenda for each DAC meeting. After NCES review and approval of the agenda, it will be shared with DAC members for their comment. Supplementary materials, such as minutes of the previous meeting, formatted so that recommendations pertinent to important decisions are emphasized, and technical papers relating to agenda items, will be distributed with the agenda. DAC members, NCES staff, and other participants will receive copies of the agenda and accompanying materials in a briefing book at least one week prior to the meeting.

Audio recordings and written minutes will be made of all deliberations at regular DAC meetings. The minutes, which will include a summary of discussion and activities that took place, will be drafted and provided to NCES staff within six weeks of each meeting. As was mentioned above, the minutes will highlight recommendations and resolutions pertinent to major NAEP issues and decisions to be made. They will also identify any actions that ETS staff will take to respond to matters that were raised but not resolved. Following NCES review, the minutes will be distributed to DAC members and invited guests within eight weeks of the meeting.

ETS will also arrange for specific work products external to scheduled DAC meetings. For example, DAC members might be asked to review and comment on a particular draft report, or to write a special paper.

NCES staff members will be supplied with copies of all correspondence and other materials exchanged with DAC members at the same time the materials are distributed to the members.

Key staff members working to integrate the deliberations and recommendations of the DAC into the overall NAEP scaling and analysis procedures will include

To continue the high caliber of advice the DAC has provided in past agreements, ETS proposes to form the core of the DAC for this agreement from current members. Should new members be required to fill any vacant seats, ETS staff will assemble the credentials of potential candidates and forward them to the NCES project officer for review. Credentials will include institutional affiliation and a biographical sketch highlighting expertise and prior work relevant to NAEP. Following comments from NCES staff, ETS will contact nominees to solicit and finalize their participation.

Current Members of DAC

It is anticipated that most, if not all, of the current DAC members will continue to serve on the committee during the period covered by this cooperative agreement. Dr. Sylvia Johnson, current chair, will be leaving the DAC to serve NAEP in another capacity. Dr. Anthony Nitko will serve as the new chair and Dr. Ingram Olkin will continue as vice chair.

Albert Beaton is professor of education at Boston College and was the director of data analysis for NAEP from 1983 to 1990. Dr. Beaton currently directs the analysis activities for the Third International Mathematics and Science Study (TIMSS). Among other contributions, he has developed procedures for item sampling (BIB spiraling) and for scaling (the ARM method). He chairs the technical advisory committee of the International Education Association.

Johnny Blair is associate director of the Survey Research Center, University of Maryland. Over the past 25 years, he has been responsible for the design and implementation of scores of surveys utilizing a wide range of sampling and data collection methodologies. Currently, he serves on the Editorial Board of *Public Opinion Quarterly*. He has published and presented papers in several journals. A recent publication is the book, *Designing Surveys: A Guide to Decisions and Procedures*, co-authored with Ronald Czaja.

Jeremy Finn is professor of education at State University of New York Buffalo. He has published extensively on multivariate analysis and statistical inference. He has written a large number of reports and articles on current educational problems, two books on multivariate analysis, and a widely used computer program for multivariate analysis (MULTIVARIANCE). Dr. Finn has a continuing interest in school and classroom processes and has been an

American Statistical Association/National Science Foundation fellow at NCES.

Paul Holland is professor of education and statistics at the University of California, Berkeley. He has written widely in the areas of statistics and psychometrics, with publications in a variety of journals. He has also authored and edited several books, and served as associate editor of the *Journal of the American Statistical Association* and *Sociological Methodology*, a member of the managing committee of the *Journal of Educational and Behavioral Statistics*, president of the Psychometric Society, and fellow at the Center for Advanced Study in the Behavioral Sciences.

Huynh Huynh is professor of education at the University of South Carolina. He has broad interests in measurement and statistics, including mastery testing and criterion-referenced testing, cut-score determination, test reliability, validity, and problems of research design. Dr. Huynh has published numerous papers and journal articles on theoretical psychometric and statistical problems.

Sylvia Johnson, who has served as chair of the DAC for many years, will be serving in a different NAEP advisory capacity, as a member of the Advisory Committee on Education Statistics (ACES). Her tenure on DAC will conclude at the end of the 1997-98 fiscal year. A replacement for her will be selected using the procedures outlined earlier.

Edward W. Kifer is professor of educational policy studies and evaluation and director of the Office of Educational Research and Graduate Studies at the University of Kentucky. He was chair of the International Technical Committee for the Second International Mathematics Study, a Spencer Foundation post-doctoral fellow at the International Association for the Evaluation of Educational Achievement, and American Educational Research Association senior research fellow at NCES, where he completed a

report on using NAEP data. He has published on topics such as attitude measurement, the Rasch model, and performance assessment.

David Lohman is professor and chair of the division of psychological and quantitative foundations at the University of Iowa, where he teaches courses in educational psychology, cognitive psychology, and the psychology of individual differences. His research interests concern the development of psychological theories of ability and achievement and the theory-based tests of these constructs. Many of his publications on these topics can be found in chapters in *Educational Measurement* and *Test Theory for a New Generation of Tests*, and in articles in the *Review of Educational Research* and the *Educational Researcher*.

Serge Madhere is professor of psychology at Howard University and codirector of the Talent Development Middle School Program at the Center for Research on the Education of Students Placed at Risk. His research interests and publications have centered on the measurement of intelligence, academic achievement and grade retention, self-esteem, violence prevention, and values and life orientations. In the past 10 years, his work has been twice recognized for excellence by AERA.

Anthony Nitko is professor of education at the University of Pittsburgh, where he teaches research methodology and measurement in the Department of Educational Psychology. His research interests include curriculum-based, criterion-referenced testing; integration of testing and instruction; classroom assessment; and the assessment of knowledge and higher-order thinking skills. He is author of many books and articles on these topics. He was elected fellow to the American Psychological Association and has served as president of the National Council on Measurement in Education.

Ingram Olkin is professor of statistics and education at Stanford University. Dr. Olkin is internationally known in the area of statistical models in the social and behavioral sciences. He has published numerous books and articles on statistics, particularly in the area of multivariate analysis. Dr. Olkin was an American Statistical Association/National Science Foundation fellow at NCES and was a candidate for the presidency of the American Statistical Association. Dr. Olkin currently serves as vice chair of the DAC.

Tej Pandey is a consultant for the California Assessment Program for the California Department of Education. He specializes in the use of psychometrics in assessment-related problem areas. Dr. Pandey was a member of the National Council of Teachers of Mathematics writing team for the evaluation section of the *Curriculum and Evaluation Standards for School Mathematics*, as well as a member of the NCES Technical Review Panel.

Hariharan Swaminathan is professor of education and psychology at the University of Massachusetts. He has also served as associate dean of academic affairs and acting dean of the School of Education at that university, president of the Educational Statisticians Special Interest Group of AERA, member of the Graduate Records Examination Board, and member of several editorial boards and national and international policy boards. Dr. Swaminathan has written extensively on topics in statistics and psychometrics and has coauthored two widely used books on item response theory.

Extra efforts will be made to recruit women to serve as members of the DAC. In the past, as many as four women have served concurrently as members. Unfortunately, the many requests women in the assessment community receive to serve this type of role often make it difficult to recruit and retain women members.

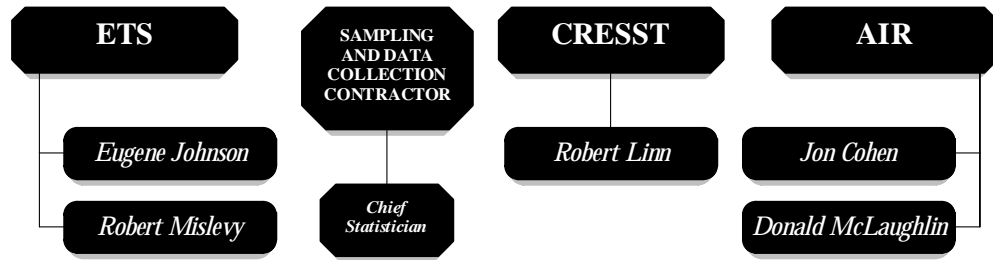
Technical Advisory Group

A technical advisory group (TAG), consisting of high-level psychometricians, statisticians, and policy analysts from ETS, AIR, CRESST, and the organization charged with sampling and data collection for NAEP, will convene to address specific urgent planning or problem-solving issues. This group will meet via teleconferences or, if necessary, in person.

The complexities of the NAEP program, combined with pressures for rapid production of reports and analyses, often lead to questions that require immediate answers. Some recent examples have been questions about the appropriateness of scaling weights derived in a multiple-sampling situation, and the representativeness of estimates of the performance of Asian and Pacific Island students. In such circumstances, decisions must be made before the analyses that are under way can proceed. A vehicle for immediately holding a thorough expert discussion of such issues and of determining necessary next steps is imperative. However, given the emphasis on reporting NAEP results quickly, there is little time for these activities, and the use of an established committee of individuals familiar with the entire NAEP project and its methodology is proposed as the most viable means of making such decisions. Fortunately, the new NAEP partnership allows us to form a senior-level advisory group familiar with NAEP and having unparalleled expertise in statistics and psychometrics. ETS, AIR, and CRESST each offers some of the nation's leading psychometricians for service on this new technical advisory group (TAG). We therefore propose that such a group be added to the NAEP program.

Meetings of this committee will be conducted on an as-needed basis to provide a forum for discussion of time-sensitive issues.

CHAPTER 2. CONDUCT TECHNICAL REVIEW GROUP MEETINGS



Deliverables

The deliverables for this task will include, for each DAC meeting, an agenda, a briefing book, and a summary of the meeting in the form of minutes. Specifics of these deliverables are described above. For the TAG meetings the deliverables will be memoranda, as appropriate, discussing issues and decisions.

CHAPTER 3.

PREPARE BRIEFING MATERIALS

Task 4

EXECUTIVE SUMMARY

Objective: To assist the National Center for Education Statistics (NCES) in providing interested individuals and organizations with timely information about the nature, findings, and progress of the NAEP project. To meet this objective, Educational Testing Service (ETS) and American Institutes for Research (AIR) will:

- prepare for the Department of Education's Office of Educational Research and Improvement (OERI) up to 24 sets of briefing materials per year, throughout the life of the project,
- submit draft briefing materials to the NCES project officer within two weeks after the request for such materials is made, and
- at NCES's direction, attend organizational meetings and professional conferences and provide up to 30 briefings about the NAEP project over the life of the cooperative agreement.

TASK 4. PREPARE BRIEFING MATERIALS

NAEP is our nation's leading achievement indicator. Its high technical quality and its political independence have placed it at the forefront of our nation's educational information system. Given its preeminence, it is reasonable to expect that the National Center for Education Statistics (NCES) will be called upon frequently by Congress, other offices within the U.S. Department of Education, or other federal agencies to provide information on the nature of NAEP, the progress of the project, and its findings. Moreover, a large number of professional organizations—such as the American Educational Research Association, the Council of Chief State School Officers, the National Council of Teachers of Mathematics, and the National Council of Teachers of English—have been involved with the NAEP project and have expressed a continuing interest in its findings.

In order to assist NCES in providing NAEP briefings to government agencies and professional groups, and to generally aid in the dissemination of NAEP results to interested professional communities, Educational Testing Service (ETS) and American Institutes for Research (AIR) will provide briefing materials to NCES at their request and, where appropriate, conduct briefings and make presentations to interested agencies and organizations.

Briefing Materials

ETS and AIR will prepare and provide to NCES up to sets of briefing materials per year (i.e., up to sets of materials over the life of the cooperative agreement). To provide a depth of staffing that ensures timely response to any and all requests, both organizations will be involved in these activities.

Briefing materials will include:

- memoranda on technical, policy, or budgetary issues,

- tabular or graphical displays of NAEP results,
- transparencies or MS Powerpoint files, and
- foam board displays (like those used at press conferences).

The specific nature of the briefing materials will be decided upon in consultation with the NCES project officer, who will initiate a request for such materials. Within two weeks of such a request, ETS or AIR will deliver draft materials to the project officer for review and comment. Within one week of receipt of comments, ETS or AIR will deliver final materials.

Present Findings at Annual Meetings

One of the most effective ways to disseminate NAEP results is through presentations at annual meetings of professional organizations. During ETS's stewardship of NAEP, its test development and statistical analysis staff have regularly presented NAEP findings and NAEP-related research at the annual meetings of the American Educational Research Association, the American Statistical Association, the Psychometric Society, the National Council on Measurement in Education, the National Council of Teachers of Mathematics, the National Council of Teachers of English, the International Reading Association, and the Council of Chief State School Officers. AIR and its staff have also maintained a strong presence at these conferences, presenting findings from the evaluation activities, Technical Review Panel and National Validity Studies Panel research activities, and research conducted under the auspices of the Education Statistics Services Institute (ESSI). Very often, ETS and AIR presentations have been part of

symposia or paper sessions in which NCES staff have also been active participants.

ETS and AIR assume that such activities will continue throughout the period covered by this Application for Cooperative Agreement. Therefore, we are planning to prepare and conduct up to 30 briefings over the life of the agreement. In all cases, we will inform the NCES project officer of the nature and schedule for these presentations well before they occur. We anticipate that most of these briefings will occur at the annual meetings of professional organizations. However, we are also prepared, at NCES's request, to conduct such meetings for Congress or other federal agencies.

CHAPTER 4.

COORDINATION WITH OTHER NAEP GRANTEES AND CONTRACTORS

Task 5

EXECUTIVE SUMMARY

Objectives: To ensure effective and regular communications among the National Center for Education Statistics (NCES), its scoring, analysis, and reporting grantees, its sampling and data collection grantee, and other NCES and National Assessment Governing Board (NAGB) contractors; and to provide prompt and accurate responses to other NCES-authorized requests for special data analyses. To accomplish these objectives, Educational Testing Service (ETS) and its subcontractors will:

- continue their current practice of conducting planning meetings with NCES that will include representatives from ETS's subcontractors as well as NCES's NAEP data collection grantee,
- attend field staff training sessions conducted by the data collection grantee,
- provide materials or support requested by the National Academy of Sciences, or whatever agency assumes responsibility for ongoing evaluations of NAEP, and
- work on NCES-authorized special analysis requests for other NAEP data users.

TASK 5.

COORDINATION WITH OTHER NAEP CONTRACTORS

This chapter has two main sections. First, we discuss mechanisms through which we will maintain open and effective project communications. Second, we will discuss our planning for the special analyses needed by many users of NAEP data who are sanctioned or supported by the National Center for Education Statistics.

Maintaining Effective Project Communication

The governance and management of the NAEP project is multifaceted. Its success demands timely and effective communication among the set of organizations that collectively set policy for, manage, and carry out the project. In particular, accurate and current information about program plans, goals, and procedures must be maintained among NCES, its NAEP data collection grantee, its NAEP scoring, analysis, and reporting grantee, and other NCES and National Assessment Governing Board (NAGB) contractors. Effective communication will be especially important during the 2000, 2001, and 2002 assessment cycles as the program begins to phase in the changes to procedures, analyses, and reports called for by NAGB.

In the past, regular and ongoing communication with NCES staff has proven essential. These regular communications have effectively ensured that all participants are kept abreast of key program decisions, and have enhanced the feelings of collegiality among the staff of NCES and its NAEP grantees.

To monitor adherence to the plans related to the 2000, 2001, and 2002 assessments and to ensure that any changes are fully

understood by all parties, Educational Testing Service (ETS) proposes to continue its practice of convening quarterly contractors' meetings among key staff members from NCES and its NAEP scoring, analysis, and reporting grantees (ETS, National Computer Systems (NCS), American Institutes for Research (AIR), the Center for Research on Evaluation, Standards, and Student Teaching (CRESST), and Aspen Systems). ETS further proposes that representatives of the recipient of the data collection cooperative agreement attend these meetings. Furthermore, at NCES's request, ETS will provide for the attendance of staff from other NAEP contractors or from NAGB.

contractors' meetings will be convened by ETS, which will coordinate meeting dates with the calendars of those involved, develop agenda, guarantee the availability of appropriate meeting materials, notify participants, and make the necessary meeting arrangements.

Coordinating operational activities among ETS, its subcontractors, and the data collection grantee is essential to

success of the NAEP project. While much of this coordination occurs at contractors' meetings, ETS has found that additional coordination activities among its staff, its subcontractors, and the data collection grantee must occur regularly throughout the project.

In addition to ensuring effective communication among the NAEP grantees, it is important that the contractor's project staff be available to NCES as a resource to support their activities and, where necessary, represent the program, at the meetings of groups whose deliberations are vital to the success of NAEP. Attendance at the meetings of the Advisory Council on Education Statistics (ACES) and the National Validity Studies Panel (NVS) have proved necessary and useful from time to time. We propose having up to three staff members (one each from ETS, AIR, and CRESST) attend the quarterly meetings of organizations such as ACES and the NVS Panel.

Conducting Analyses for NCES-Supported NAEP Data Users

In addition to attending regular meetings, ETS proposes to support NCES grantees and other users of NAEP data through the provision of special analyses. Since its inception, NAEP has been recognized as the nation's leading academic achievement indicator. The increasing recognition of the state-of-the-art technical quality of the program, the addition of state-by-state assessments in 1990, the emergence of a standards-based reform movement, and NAGB's decision to establish and report NAEP results in terms of achievement levels have further increased the interest in, importance of, and visibility of NAEP results. With this increase in prominence has come an increased demand for NAEP results by other federal and state government agencies, the media, educational researchers, and other educationally-oriented organizations. The information needs of some of these organizations are met by ETS-developed secondary-user products, such as restricted-use data files and related software and Web-based summary data tables. However, with increasing frequency NCES is confronted by organizations with data needs that require ETS to conduct special-purpose analyses of NAEP data.

For example, over the last two assessment cycles special analyses of NAEP results have been authorized by NCES and conducted by ETS in support of the National Education Goals Panel annual reports. The National Council of Teachers of Mathematics has published a special volume on the nation's mathematics achievement that is based largely on item-level results from recent NAEP assessments provided to them by ETS at NCES's behest. More recently, *Education Week's* 1998 publication of *The Urban Challenge: Public Education in the 50 States* made extensive use of state-by-state analyses of academic achievement data from the nation's urban public schools. These analyses were conducted by ETS expressly to support this report.

Areas of interest requiring technical assistance include: special sample extraction, analyses supporting achievement-level setting, special summary tables, ad hoc analyses, test item information, special scaling studies, information retrieval, and special reporting requests.

ETS and its subcontractors fully expect the frequency of requests for special analyses to increase. As continuing to enhance the utility of NAEP data is an important priority for both NCES and NAGB, ETS understands the need for prompt and accurate response to such requests. To meet the burgeoning demand for special analyses, ETS and its subcontractors AIR and NCS will, in combination,

advise on, carry out, and deliver NCES-authorized requests. Of course, in practice we would expect a large number of specific individuals to be involved in analysis tasks depending on the nature and level of expertise required.

Moreover, having three organizations, AIR, ETS, and NCS, committed to filling these requests will facilitate rapid completion.

CHAPTER 5.

PROVIDE SUPPORT TO NCES, NATIONAL ARCHIVES, AND NAGB AND PROVIDE MONTHLY REPORTS

Tasks 6, 7, and 8

EXECUTIVE SUMMARY

Objective: To provide appropriate analytic and logistical support to the National Center for Education Statistics (NCES), the National Archives, and the National Assessment Governing Board (NAGB). To accomplish these goals, Educational Testing Service (ETS) and its subcontractors will:

- continue to provide NCES with progress reports which provide task-by-task descriptions of work in progress, status of deliverables, and current expenditures;
- plan for project staff to attend meetings of the National Assessment Governing Board (NAGB);
- provide materials or support requested by the National Archives; and
- provide special analyses requested by NCES and NAGB.

INTRODUCTION

Educational Testing Service (ETS) and its subcontractors are committed to providing the most accurate and timely service to the National Center for Education Statistics (NCES), the National Archives, and the National Assessment Governing Board (NAGB). NCES and NAGB are the two organizations charged with leading NAEP, and operational grantees provide a range of support to these organizations, and to the other groups with whom they work.

Support to NCES and NAGB will take three major forms: The first is the regular and accurate provision of information necessary to allow NCES and NAGB to set policy for the program and to manage its ongoing operations. The second type of support involves special analyses that support the work of NCES and NAGB, or of other organizations such as the National Education Goals Panel. Third, ETS and its partners are committed to being available to give NCES and NAGB up-to-the-minute information on any matters related to NAEP.

TASK 6.**PROVIDE SUPPORT TO NCES
AND THE NATIONAL ARCHIVES**

ETS and its subcontractors stand ready to support NCES in a wide variety of ways. Much of the strength that the ETS team brings to NAEP is our proven flexibility: We have shown the ability to handle a range of short-term assignments and to deal promptly with unforeseen issues. As questions or issues arise in NAEP, ETS will provide whatever support NCES believes is necessary. Such support may involve one or more of the following activities:

- organizing or participating in special-purpose conferences NCES wishes to conduct,
- organizing or participating in meetings designed to discuss or resolve technical or logistical issues,
- preparing memoranda, papers, or other materials in support of such meetings or conferences,
- writing position papers that recommend courses of action on NAEP operational or logistical matters,
- conducting other special analyses—either of NAEP data or of operational or budgetary matters—requested by NCES, and
- working in support of groups with whom NCES is collaborating, including the National Education Goals Panel (NEGP), the Advisory Council on Education Statistics (ACES), and state departments of education.

Many of these tasks overlap with those discussed in Chapter 4, and all of the text will not be repeated here.

Our goal in these tasks is to facilitate NCES's tasks in managing the implementation of NAEP. We therefore stand ready to respond to requests for information and support in a timely fashion.

NAEP is a data-collection and analysis activity conducted in the public interest and at public expense. It is therefore essential that all NAEP information be provided to the main repository of national information, the National Archives. ETS will, when secondary-use data files are delivered to NCES, also provide copies to the National Archives. The files will be delivered in an ASCII format. In addition, ETS will provide an electronic copy of the user guide that accompanies the data files. The scope of work statement in the Application for Cooperative Agreement specifies that the user guide also be delivered in ASCII format. ETS will, if instructed to do so, deliver the manual in that format. However, doing so would risk losing all the document layout and formatting. We have noted, in regulation 36 CFR 1228 (in the *Federal Register* of October 21, 1997) that the National Archives will accept textual documents with Standard Generalized Markup Language (SGML) tags. Therefore, we propose to work with the NCES NAEP project officer and representatives of the National Archives to determine the format in which user guides should best be submitted.

TASK 7.

PROVIDE ADMINISTRATIVE REPORTS

Supporting NCES involves not just special analyses. Accurate and clear project management information is also necessary. Each , ETS will submit progress reports to NCES providing an update on the status of the project. These reports, which will be submitted electronically will be in two parts: The first part will be a project report including a description of the project activities . T

his project report will discuss major accomplishments, significant findings, and problems in need of resolution. The second part will be a financial report comparing budgeted and actual expenses for the period to date and will include a projection of future costs of completing the various tasks.

Within one month of the award of the cooperative agreement, we will meet with NCES staff to review the proposed contents, layout, and delivery schedule and to make whatever revisions NCES would like to increase the utility of these reports. We will also work with NCES to determine the appropriate grouping of individual project tasks to furnish the most useful and effective project status on a monthly basis.

TASK 8.

PROVIDE SUPPORT TO THE NATIONAL ASSESSMENT GOVERNING BOARD

As the body in charge of setting policy for NAEP, NAGB plays a central role in the program. It is therefore necessary that those charged with implementing the operational NAEP program be familiar with the deliberations, actions, and policies of the Board. It is also important that we support the Board in any ways indicated by NCES. Such support may involve presentations at NAGB meetings, special analyses in response to NAGB requests, and policy papers and memoranda. ETS and its subcontractors propose to meet the support needs of NAGB in a variety of ways.

First, the attendance of contractor staff at NAGB meetings has been an important aid to NCES in providing information to the Board, as well as keeping the contractors' staff informed of the Board's policies and program concerns. We propose that staff members

attend NAGB meetings. Of course, the number of staff members attending from each organization will vary; for some NAGB meetings, it may be unnecessary that staff members from some of the partner organizations attend. We will ensure that no unnecessary trips to NAGB meetings are made.

Second, we will be prepared to make any presentations at NAGB meetings that NCES requests. Along with providing staff to make presentations, we will prepare print materials and transparencies to support these presentations. These materials will

be submitted at least one month prior to NAGB meetings so that they may be included in NAGB briefing books. While the complexity and number of such presentations may vary, we will assume for planning and budgeting purposes that ETS staff will make two such presentations at each NAGB meeting.

Third, we will provide any materials requested by NCES to support presentations at Board meetings made by NCES or NAGB staff. These materials will also be submitted in time for inclusion in briefing books. While the complexity and number of these presentations may vary, we will assume for planning purposes that two such sets of support materials will be needed for each NAGB meeting.

Fourth, if requested by NCES, we will conduct special analyses and write reports, papers, or memoranda needed by NAGB. Such documents will be submitted to NCES for review prior to their submission to NAGB. While it is difficult to predict the number of such requests, we will assume for planning purposes that we will undertake six such tasks at NAGB's request per year.

Fifth, ETS and National Computer Systems (NCS) will gather and copy, at NCES instruction, all materials needed by NAGB contractors in support of activities such as achievement-level investigations, or other analytic matters. However, we should note that because all subjects covered under this procurement will already have had achievement levels set, we have assumed lower levels of activity than in the past. For the same reason, we have not included budgetary provision for attendance at meetings of the Technical Advisory Group on Standard Setting. Attendance at these meetings for current standard-setting projects is covered under the extant cooperative agreement.

In all cases, ETS and its subcontractors will provide materials to NAGB only when instructed to do so by NCES.

CHAPTER 6.

PROVIDE SUPPORT TO STATES

Task 9

EXECUTIVE SUMMARY

Objective: To maintain the close partnership with the states that has characterized the state NAEP program and to expand upon the services provided to the states. To accomplish this objective, Educational Testing Service (ETS) and its subcontractors, American Institutes for Research (AIR) and Center for Research on Evaluation, Standards, and Student Testing (CRESST), will:

- provide *general technical assistance* and consulting services, available through the NAEP/ETS toll-free number,
- publish *The State of Education*, to provide formal National Center for Education Statistics (NCES) communications about the structure and schedule of the state assessment program,
- conduct regular NETWORK meetings in Washington, DC, at strategic points in the state assessment cycle,
- produce *NETWORK News*, to provide timely informal advisories to the states about the conduct and progress of the state assessments,

- provide consultation and assistance to states interested in linking their state assessments to NAEP, and
- hold state NAEP conferences, where states can report on and learn about ongoing research activities pertaining to the state program.

TASK 9.

PROVIDE SUPPORT TO STATES

From the inception of state NAEP, the states have been active participants in program design and implementation. State representatives have contributed to the development of curriculum frameworks; developed and reviewed assessment exercises and background questionnaires; played a key role in the recruitment of schools; helped formulate program policies for sampling, data collection, and reporting; and assisted in the design of reports. The participation of the states has been central to the success of the national assessment and its state component.

Throughout its tenure on NAEP, Educational Testing Service (ETS) has maintained a close and effective partnership with the states, and provided a range of services to them. These services have included general technical assistance, formal presentations at state and national conferences, and ad hoc assistance to state policymakers, state education officials, and others interested in NAEP. Moreover, ETS has conducted regular meetings with representatives from state education departments (NAEP NETWORK meetings) and produced regular publications and newsletters (*The State of Education* and *NAEP NETWORK News*), as a forum for state input on program issues and timely dissemination of program news.

ETS and its subcontractors, American Institutes for Research (AIR), Aspen Systems, and the Center for Research on Evaluation, Standards, and Student Testing (CRESST), plan to maintain the close partnership with the states that has characterized the state NAEP program. To this end we will continue to:

- provide general technical assistance and consulting services, available through the NAEP/ETS toll-free number,

- publish *The State of Education*, to provide formal NCES communications about the structure and schedule of the state assessment program before each state assessment,
- conduct regular NETWORK meetings in Washington, DC, at strategic points in the state assessment cycle, and
- produce *NETWORK News*, to provide timely informal advisories to the states about conduct and progress of the state assessments.

The NAGB *Policy Statement on Redesigning the National Assessment of Educational Progress* has as its third core objective, “To help states and others link their assessments with the national assessment and use national assessment data to improve performance.” To meet this objective, ETS, AIR, and CRESST will expand our support to the states by providing two new services:

- assistance to states in the projection of state assessment scores onto the NAEP scale, and
- program-sponsored state NAEP conferences, where states can report on and learn about ongoing research activities pertaining to the linking of large-scale assessments and the use of NAEP data to improve instruction, education, and assessment in their states.

Each of these areas is discussed in more detail below.

General Technical Assistance

ETS has routinely provided considerable technical assistance to state policymakers, state education officials, and others interested in NAEP. We plan to continue this type of assistance under the current procurement. The content and coverage of this technical assistance will include:

- consulting with researchers who are pursuing investigations using state NAEP results and conducting secondary analyses of NAEP data,
- providing information to state NAEP data users on the procedures used to draw the samples, administer the assessment, score, and analyze the data,
- providing curriculum planners and policy analysts with subject-specific findings on student performance, instructional experiences, and home and school supports for learning,
- supplying background information to state and local test administrators and accountability committees about the development and nature of assessment instruments, and
- assisting states in accessing and using data from NCES's Web site.

ETS will also continue its provision of a toll-free number to facilitate the processing of state technical assistance requests.

The State of Education

Since the inception of the state assessment program, NCES has used *The State of Education* as a means of formally communicating with the states about the structure and schedule of the state assessment program. Prior to each assessment cycle, ETS has produced for NCES an issue of *The State of Education* to formally announce program offerings (subjects and grades to be assessed); assessment content (an overview of the frameworks); procedures for state review of assessment materials; information about achievement levels; state review procedures for reports; and program standards for participation, test administration, test security, data collection, and reporting.

ETS will work with NCES to develop two issues of *The State of Education*. The issues will be produced in a similar format and

will contain information similar to the editions produced for the 1996 and 1998 assessments. The first issue, which will announce the 2000 state assessment, will be developed and printed in the spring of 1999. The second issue, which will announce the 2002 state assessments, will be produced in the spring of 2001. Both issues will be available in print and will be posted to NCES's Web site.

NETWORK Meetings

The NAEP NETWORK (composed of the state testing directors, state NAEP coordinators, representatives from the Council of Chief State School Officers (CCSSO) and private school organizations, NCES, and its NAEP grantees) has remained an important institution, helping to form appropriate policies, develop high-quality assessments, and facilitate changes to reporting strategies to better serve state audiences. One of the principal forums for regular and effective communication among the states, NCES, and its grantees has been NETWORK meetings.

The meetings in the calendar year prior to a state assessment have traditionally been used as a vehicle for state subject-area coordinators to review and provide input on test items

for the upcoming assessments.

Two such meetings would occur during the period covered by the current procurement

As has been prior practice, ETS will develop an agenda in consultation with NCES several months in advance of each meeting. These agendas will be shared with members through mail, broadcast fax, and e-mail, well in advance of the scheduled meeting.

ETS will redouble its efforts under the current contract to renew and revitalize the NETWORK meetings. Changes in structure, content, and format will be worked out with NCES, in consultation with current NETWORK members as well as members of the assessment subcommittee of EIAC. Some possible changes to the NETWORK meetings and activities include:

- organizing NAEP special interest groups around issues such as report design, background questionnaires, or improvement of school participation,
- instituting periodic teleconferences of small groups focused on particular issues (e.g., a proposed change in program procedures or a suggested revision to report designs),
- adding regular features to meetings (e.g., an EIAC update), and
- adding timely update items on non-NAEP NCES activities (e.g., the Schools and Staffing Survey).

As has been the case under the existing cooperative agreement, Aspen Systems will be responsible for meeting arrangements, production and distribution of meeting materials, and the provision of on-site assistance to attendees.

NETWORK News

ETS proposes to continue its production and dissemination of the *NETWORK News*.

As necessary, additional editions on pressing special topics will be produced.

We will continue our redesign efforts to make the newsletter more useful and accessible to NETWORK members. Some possible enhancements to the newsletter itself include:

- inclusion of regular surveys of specific issues members would like to see addressed at NETWORK meetings,
- regular notices of publication of NAEP reports and research papers, and
- continuing notification of new or updated Web sites of interest to state education departments.

ETS also proposes making each issue of the *NETWORK News* available on the Web through NCES's NAEP home page.

Assistance to States in the Projection of State Assessment Scores onto the NAEP Scale

One way in which NAEP can be made most useful to state education agencies is by providing a benchmark against which they can compare the results of the census assessments they carry out in their schools. If their state assessment scores can be mapped onto

the NAEP scale, then they are in a position to make stronger statements about the implications of performance on their state assessment than would otherwise be possible. If the projections are valid, schools and districts can compare their scores not only to the state as a whole but also to the entire nation, and based on linkage of NAEP to international assessments, to other major countries in the world.

Because this capability is valuable, states have examined alternative methods for developing projections, including special linkage studies and embedding released NAEP items in their state assessment administrations. However, these methods are flawed because they do not link the operational state assessment to the operational NAEP, on which the scales are defined. Varying the context in which an item, or a test, is administered varies the performance of students on that item, so the only valid linkage must be based on linking scores on actual administrations of NAEP and state assessments.

Building on earlier work of Linn (1993), Bloxom et al. (1995), and Williams et al. (1995), Donald McLaughlin (1998a) of AIR has explored the feasibility and validity of projections based on linking state assessment scores of students onto NAEP performance records. Using the 1996 state NAEP grade 4 and 8 mathematics assessments in four states, he found (a) that it is feasible to develop the linkage of student records without violating either NAEP or state assessment confidentiality assurances, and (b) that, in three of the four states, acceptably accurate projections of mean NAEP scores and percentages at achievement levels could be obtained for sufficiently large groups of students.

Among other things, McLaughlin (1998b) found that to be neutral (i.e., so that comparisons based on projected NAEP scores lead to the same conclusions as comparisons based on actual NAEP scores), it was necessary to include (a) explicit terms for school

mean scores, as well as individual student scores, and (b) explicit terms for demographic measures. Like others (Linn & Kiplinger, 1995; Shepard, 1997), he also found that projection functions did not generalize across years.

As a part of the exploration of linkage, McLaughlin developed detailed guidelines for validating score projections and for conveying the appropriate level of uncertainty in statements based on projections. McLaughlin's methodology for linking and for development and validation of projections will be available to states,

That consultation can cover both the development and evaluation of the linkage database and the execution of a specified series of analyses to derive the appropriate projection function and evaluate the precision, neutrality, and stability of projections.

Although the scope of the NAEP project does not include funds to construct individual-level linkages of state assessments in participating states, several activities can be carried out which will enable and support the development of such linkages. In particular, linkages can be constructed based on school mean data that are sufficiently similar to more precise linkages using individual student data to provide diagnostic information to states participating NAEP about the potential for linking their state assessments to NAEP. For example, we can inform states concerning the extent to which their assessments are related to NAEP subscales and, by comparisons with earlier linkages, we can assess the extent to which their linkages vary over time.

Five linkage activities are described below. We propose to carry out these activities within the NAEP project. Additional activities supporting individual state linkage development and evaluation would be undertaken with funding outside the NAEP cooperative agreement.

School-Level Linkages in Reading for States That Have Available School-Level Assessment Data

This activity will provide a baseline of information to states participating in state NAEP concerning the statistical relations between their state reading assessments and NAEP. It will be implemented in three steps, each of which has previously been carried out to study state assessment linkages for the 1994 and 1996 state NAEP assessments. The effort required for this activity will be diminished by the fact that linkages will already have been constructed for eight of the states participating in state NAEP in 1998 as a part of an Education Statistics Services Institute (ESSI) project to study linkages between NAEP and state assessments.

Step 1 involves acquiring school-level state assessment data from each state. Most states make public the results of state assessments in their public schools. In some states these publications are at the district level, but in many they are at the school level. The process of acquiring the data varies from state to state but is normally straightforward. An announcement of this activity at a NAEP NETWORK meeting and a follow-up contact should be sufficient to obtain either diskettes, CDs, Web site addresses, or printed reports with state assessment data. Each state has a somewhat different method of dissemination, and we will take responsibility for translating from the format in which they customarily provide the data to a common format for linking to state NAEP.

Step 2 involves matching schools to construct a file containing state assessment information for each school participating in state NAEP. In most states, public schools have state-assigned identification numbers. Because these numbers are present on the Common Core of Data file, along with the NCES identification numbers which can be linked to NAEP, matching of most public schools will be very easy. In a few states, matches on school and district names and addresses may be necessary, however.

Step 3 consists of carrying out analyses to construct and evaluate school-level linkages. Using software developed in the study of linkages of state NAEP to four state mathematics assessments in 1996, we will construct projections of state assessment school means onto the NAEP scale and evaluate these projections in terms of neutrality and precision. To maximize neutrality of the projections with respect to subpopulation variations, NAEP background data will be used in constructing these linkages. These results can be used by states both to compare performance of large sets of schools on the NAEP scale and to evaluate the potential value of constructing student-level linkages between their state assessments and NAEP.

A Report Comparing School-Level Reading Linkages Between State Assessments and State NAEP

One of the crucial issues concerning linkages between tests is whether they remain constant over time. By comparing the results of the above activity to school-level linkages constructed for the 1994 NAEP reading assessment (Wu, Royal, & McLaughlin, 1997), we can assess the extent to which these linkages changed between 1994 and 1998. In many states, changes are to be expected because the state assessment instrument was changed between 1994 and 1998. In those cases, the linkage differences may be useful to states as a check on their own linkages between the old and new assessment instruments. In other states, instabilities in linkages can also be expected, because of gradual accommodation of school curricula to the state assessment objectives, compared to a more uniform level of familiarity with NAEP over time.

The report will be made available to participating states through a variety of forums, including NAEP NETWORK meetings and publication on the NAEP Web page. This report will be available to states by July 1999.

School-Level Linkages in Mathematics and Science Assessments for States That Have Available School-Level Assessment Data

This activity will provide a baseline of information to states participating in State NAEP concerning the statistical relationships between their state mathematics and science assessments and NAEP. The activity will be carried out in a manner analogous to reading linking activities described above. However, consideration will also be given to translation of state assessment results onto the new NAEP market-basket scale in mathematics.

A Report Comparing School-Level Mathematics and Science Linkages Between State Assessments and State NAEP

This activity will parallel the reading report activity and will provide the basis for a report to states by July 2001.

Provide Materials to States That Wish to Construct and Evaluate Student-Level Linkages Between State NAEP and Their State Assessments

In addition to the four analytical activities described above, ETS, CRESST, and AIR will provide consultation to states to help them construct and evaluate linkages between their state assessments and NAEP. Specifically, we will provide them with copies of guidelines for projecting scores onto the NAEP scale, which include recommendations for specific data gathering and analysis steps. We will also provide them with packaged software to use to conduct the steps required for validating a linkage. That software will be based on prototype software developed for the 1996 four-state NAEP linkage study. Finally, states will be kept informed concerning ongoing efforts to find methods for constructing more precise linkages. For example, the variant-item technique is being explored as a method for making use of the relations of individual item responses on NAEP and state assessments to produce

alternative scorings of state assessments that are optimally aligned with the NAEP scales.

State NAEP Conferences

As an important component of our effort to increase NAEP's utility to the states, we propose to organize and conduct state NAEP conferences. These conferences will fill an important gap in NAEP services to states. NETWORK meetings, while essential to the success of the program, are focused primarily on the ongoing management of NAEP. They do not provide an appropriate venue for in-depth discussion of technical and policy issues. The proposed new state NAEP conferences are intended to allow for and encourage such discussion.

State NAEP conferences will be conducted

The staff at CRESST will plan these conferences.

CRESST, AIR, and ETS will work with NCES, EIAC and state NETWORK members to select themes, topics and presenters for these conferences. We anticipate that some presenters will be individuals currently conducting state NAEP-related research. CRESST may also commission papers on key issues of interest.

We recommend that the conferences be organized around specifically defined themes so that presentations can be in-depth and, therefore, most likely to inform state practice. NCES will be given final approval for topics. However, to be consistent with the NAGB redesign policy, we recommend that the first conference be centered around the topic of linking state assessments to NAEP. If NCES concurs with this topic choice, we envision that the conference would include a summary of past research on linking tests to NAEP and give particular attention to factors that may affect

the trustworthiness of the linkage for different purposes (e.g., interpreting individual student performance, comparing performance of aggregates such as schools or districts to NAEP achievement levels, and reporting state results in reference to NAEP on years when state NAEP is not administered).

Topics for the second conference will be decided in consultation with NCES after the first conference is held. If interest in linking continues, and if states have experience to share as a result, the second conference might be a follow-up of the first with emphasis on what was learned from state linking efforts. Alternatively, another technical topic of interest to states would be selected based on advice from EIAC, NAGB, state NETWORK members, and NCES.

CRESST will provide a full range of services related to this conference. In addition to arranging for a conference location and sending invitations, CRESST will produce conference materials and arrange for presenters. CRESST will also publish or otherwise make available conference proceedings.

Invitations will be extended to states interested in participating in NAEP and other individuals interested in state NAEP matters. For planning purposes, we assume each conference will have

Consideration will also be given to scheduling these conferences in coordination with NETWORK meetings.

Other State-Support Activities

The scope of work in the Application for Cooperative Agreement suggests that “states shall have the option to use national assessment tests in other subjects (than those offered as part of the core program in any given year) and at grade 12 by assuming a larger share of the costs and adhering to requirements that protect the integrity of the . . . program.” ETS and its partners stand ready to work with NCES, NAGB, and the states to provide services if requested. These services will include printing of assessment booklets, shipping and processing, scoring, analysis, and reporting.

CHAPTER 7.

DEVELOP AND MAINTAIN STATISTICAL ANALYSIS PROGRAMS

Task 10

EXECUTIVE SUMMARY

Objective: To develop and update statistical programs and modules, as needed, to maintain functionality and compatibility with current statistical software. To provide documentation and training materials for National Center for Education Statistics (NCES) staff and other users of these modules. To achieve these objectives, Educational Testing Service (ETS) and American Institutes for Research (AIR) will:

- develop or update regression analysis and cross-tabulation modules that work with current versions of SPSS, SAS, statistical analysis programs,
- continue development of the *NAEP*EX data management and extraction program and the *AM* estimation program developed by AIR,
- conduct workshops to demonstrate all software to NCES, and
- provide all this software to secondary users at no cost.

TASK 10.

DEVELOP AND MAINTAIN STATISTICAL ANALYSIS PROGRAMS

Overview

Since the inception of the NAEP Data on Disk product in 1994, Educational Testing Service (ETS) has developed, implemented, and provided training for two interdependent software products: *NAEP**EX*, a data management and extraction tool, and analytical modules that operate within the SPSS analysis system to produce correct and consistent results within the NAEP sample design. Under a separate NAEP contract, American Institutes for Research (AIR) has developed a stand-alone software product, named *AM*, that provides estimates of subgroup effects without the use of plausible values. More detailed descriptions of these software products follow.

The continued development, enhancement, and implementation of these software products are based on ETS's and AIR's commitment to the following objectives:

- to make NAEP among the easiest of large-scale data sets to properly analyze, therefore increasing the amount of substantive and policy-relevant research questions that can be addressed with NAEP data;
- to provide users with the information and support they need to understand the powerful statistical methods used in NAEP analyses; and
- to provide secondary researchers with the tools they need to replicate all the statistical NAEP analyses conducted by the contractor.

To that end, the remainder of this chapter will outline ETS's and AIR's plans to:

- enhance, maintain, and support SAS and SPSS modules and develop similar modules for Stata and S-plus;
- introduce key improvements in *NAEPPEX*, including a more accessible user interface, the ability to extract assessment items either in their raw form or scored exactly as they were in the original analyses conducted by the contractor, and extraction of data to files directly readable by popular statistical software such as SAS or SPSS, as well as the new *AM* software; and
- continue development of the *AM* software by performing additional testing, improving the user interface, and providing on-line help. The goal will be to provide a copy of *AM* to every NAEP data user from 2000 forward.

SPSS and SAS Statistical Modules

The multistage sample design of NAEP requires analytical methods more sophisticated than those offered by current, off-the-shelf statistical packages, which assume simple random samples in their analysis procedures. ETS, in conjunction with Westat Inc., developed a jackknife-based methodology for the computation of standard errors of reported statistics. Another important feature of NAEP analysis is the use of plausible values to account for the measurement error inherent in the matrix sample design employed in NAEP.

Since 1986, every user guide accompanying the release of a NAEP secondary-use data product has contained examples of SAS and SPSS code for computing the mean and standard error of a given NAEP variable by a single categorical variable. This code is complex, involving multiple passes and aggregations of the relevant variables, and assumes a high level of user sophistication to code, modify, and perform correctly.

In 1994, ETS initiated the development of statistical procedure modules that operate under the SPSS/Windows product, use SPSS-style command syntax, and compute the standard errors correctly and consistently. The first module performed two-way cross-tabulations and ad hoc comparisons of column, row, and cell effects. Subsequent additions to

this library included a correlation and regression analysis module and a multi-way tabulation module. These modules were further enhanced by incorporating them into the SPSS/Windows menu bar, which offers drop-down forms for the collection of user specifications of variable names and processing options and produces correct command syntax for their execution.

An important consideration in the selection of the SPSS/Windows product as the initial system for analytical module development was the availability of the SPSS Developer's kit for the Windows platform and the level of support provided by SPSS to extant compiler technology. At the time that the modules were originally developed, SAS supported its Toolkit product only for mainframe, Unix, and OS/2 platforms and was developing one for the Windows platform. SAS later released the Toolkit for Windows product and ETS subsequently acquired the license for development of the NAEP analysis modules under the SAS system. The ability for ETS to develop and update modules for any statistical package on any platform is dependent on the vendor's schedule for release of the developer's kit corresponding to the new release of their analysis product.

NAEPEX

In 1994, ETS developed *NAEPEX*, a PC/DOS-based data management and extraction program, to assist NAEP secondary-data users in the manipulation of the NAEP Data on Disk CD-ROM products. The program is data independent; each Data on Disk product contains the control files necessary to define the assessment component and constituent samples to *NAEPEX*. A Windows-based version of the program was developed by ETS staff in 1996 and released in 1998.

NAEPEX will guide the user through the process of selecting samples and data variables of interest for analysis and can create an extract data file or a set of SAS or SPSS control statements, which will define the data of interest to the appropriate analysis system. The user can then execute these control statements under SAS or SPSS and proceed with

performing the required analyses. In this way, *NAEPEX* complements the NAEP statistical analysis modules by generating the data and control statements for processing by the respective systems and the special modules.

***AM* Software**

The *AM* software was originally developed with funds from the National Center for Education Statistics (NCES) to estimate single assessment scale, marginal maximum likelihood (MML) regression models, and to serve as a platform for testing and evaluating alternative statistical models. A fundamental characteristic of the platform is that it provides Taylor series estimates of all standard errors, based on Binder's (1983) general formula. A second round of NCES funding added the capacity to estimate marginal maximum likelihood regressions that took composite scales as a dependent variable, narrowing the gap between the program's functionality and the operational requirements of NAEP.

Recognizing the wide acceptance of plausible values by the research community, the AIR team working on the software obtained internal AIR funding to develop the appropriate Bayesian methods and integrated them into the *AM* software. This module is very new and has only undergone initial testing, but simulations in a prototype proved to be highly successful.

In addition to the NAEP module, the software package also includes two additional "bonus" modules that have been added and are part of the freely available software. They include a module for estimating basic descriptive statistics (means, standard errors, etc.) and a module for estimating robust least squares regression. These are the sorts of procedures that NAEP users might employ in analyzing NAEP data (e.g., background and instructional practices data). Both of these modules exploit the software's fundamental ability to provide standard errors appropriate for a complex sample design.

Designed as a statistical platform rather than a fixed set of procedures, adding new modules is simple and fast. AIR intends to add capabilities as they are requested by NAEP constituents, subject to NCES approval.

Product Development and Support

Module Development and Maintenance

For the duration of the contract, ETS will continue to develop, improve, and promote the NAEP library of SPSS and SAS procedures. ETS will also draw on the expertise of its staff to develop, implement, and support analogous analytical modules for the Stata and S-Plus systems. Where conditions permit and user interest is demonstrated, ETS will develop and support these modules for other platforms, including Unix and mainframe. In addition, ETS will release new versions of these modules within 60 days of any new commercial releases that make the modules obsolete or non-functional.

As each new module is developed and prototyped, ETS will schedule a critique/training session with appropriate NCES staff on the use and operation of the module. When a module is updated either for new releases of the analytical system or enhanced features, ETS will notify NCES and jointly determine if training is necessary.

ETS will work with NCES staff to develop a delivery system for the new and updated modules and documentation through the NCES Web site. This will ensure that the end user of the software will have immediate access to the appropriate version of the software as its release is announced.

NAEPEX Enhancements

ETS will continue to develop, support, and provide training for *NAEPEX*. Some improvements already under consideration are an improved user interface, the ability to manage more than one set of data concurrently, and

the capability for scoring cognitive items according to the scaling procedures used by ETS.

Concurrent with the development of analysis modules for *NAEPEX* will be enhanced to accommodate the transition of the raw data files. Such enhancements would include the generation of syntax for data definition and formatting of the extracted data into blank-, comma-, or tab-delimited formats for easy input to database and spreadsheet programs.

ETS will develop a Web-based version of *NAEPEX* that would permit access to the public-use data files as they become available on the NCES Internet server. This version would permit Internet users to construct and extract customized data files and related documentation and syntax files directly from the NCES Web site without having to download the entire set of files for each sample of interest.

ETS will also continue to monitor and evaluate Web tools and technologies as they evolve, with the aim of developing a Web-compliant version of *NAEPEX* that could operate in a stand-alone mode on Macintosh and Unix platforms as well as PCs. Among the advantages to this enhancement of the software are: Development costs are kept low by using standard, readily available Web tools and techniques; the software can be distributed in a stand-alone form with the restricted-use data for any computing platform that supports Web browsers; and maintenance and upgrades to the software will be directly influenced by new advances in Web technology.

AM Development, Testing, and Release Plan

Development and Testing

In the first year, ETS and AIR will compare estimates obtained through both *AM* and existing plausible values methodology, using artificial data first, then actual NAEP data. Initial comparisons will examine single-scale

regression coefficients and standard errors using simulated data sets. After resolving any differences between AM and the current plausible value methodology, we will estimate the same models using actual NAEP data for confirmation.

We then will move on to analyzing composite scales and comparing estimates of regression parameters and the covariance matrix of the residuals for the various methods, using first artificial and then actual NAEP data. We will track down the sources of any differences encountered and, in conjunction with NCES and the DAC, decide how to resolve them.

We will next introduce the Bayesian component to AM and compare the characteristics of the posterior distributions estimated from each program, again using artificial data first, then actual NAEP data. Comparisons will include full-sample percentile distributions, percents above cut scores, and standard errors. These statistics will be replicated for a variety of subgroups, including cross-classifications. Again, any differences will be identified and resolved.

In the second year, we will release a beta test version of the software to coincide with the release of the 1998 NAEP data. ETS and AIR will work jointly with NCES to identify potential beta testers of the software. An alternative procedure for testing the software is described in the section below.

We will ask beta testers to register and report any problems, comments, or suggestions regularly. We will maintain a telephone contact number as well as an e-mail address for the contact. Throughout the test period we will implement patches to the program as necessary, and keep a log of technical support questions and comments. We will design and implement changes and enhancements to the software that will avoid the most common questions or complaints.

Release of Operational NAEP 2000 Version

This phase will include the release of the official version of *AM*, along with the 2000 NAEP assessment data. The software will be released with a full users' manual, on-line documentation, and NAEP data in native *AM* format, or directly and transparently exportable to *AM* from *NAEPEX*. ETS and AIR will arrange to offer software training at professional conferences near the time of the release, and will provide telephone technical support throughout the year.

Testing of Software

ETS and AIR will endeavor to release new versions of the above software products to coincide with the first release of secondary-use data products in each assessment cycle. When possible, a pre-release copy of the software and a draft user guide will be shipped to NCES along with the adjudication version of the data for their evaluation and feedback. As discussed in Task 55 in Chapter 19 and Task 174 in Chapter 48, training sessions for state personnel will be scheduled around the time of the release of the state reports and data. In the past, these sessions have provided an excellent opportunity to introduce new software and enhancements to current software, and to receive immediate feedback on its usability and functionality.

If invited, ETS will also participate in NCES-sponsored workshops on the use of NAEP data. Previous participation in these workshops has proven to be beneficial to all parties involved: ETS staff establishes and, to a large part, maintains contact with users of the data and software; NCES gains a broader base of NAEP data users; and the attendees receive the best instruction on the complexities of the NAEP design and in the application of the software.

Upon completion of these sessions and workshops, ETS and AIR will make final modifications to the software and documentation based on feedback from those sessions. The final version of each software release will then be sent to NCES for dissemination to the user community.

CHAPTER 8.

DEVELOP SETS OF BACKGROUND QUESTIONNAIRES

Task 11

Objective: To develop streamlined and effective sets of background questionnaires. To accomplish this objective, American Institutes for Research (AIR) and Educational Testing Service (ETS) will:

- construct a new background questionnaire framework and use this framework as a guide for development,
- review existing questionnaires to identify effective items,
- implement a modular design composed of core modules to be administered with each assessment and non-core modules to be administered either one time or at intervals of several years,
- develop the core and non-core modules in a way that improves utility and reduces respondent burden,
- use cognitive laboratories in the development process, and
- develop and implement a strategy for bridging between old and new versions of background questions.

INTRODUCTION

Background questionnaires are a key part of the NAEP program. Over the years, NAEP has gathered information on instructional practices and school and classroom policies that has provided an important context for the reporting of achievement-level results. However, this data collection has come at the price of examinee burden. For NAEP in 2000, 2001, and 2002, we propose, under the leadership of survey experts at American Institutes for Research (AIR), to design and craft a streamlined and more informative set of background questionnaires.

In this chapter, we present an overview of the procedures that we will follow to create shorter, but improved, background questionnaires that will support accurate estimation of the relationships between educational achievement, classroom practices, and students' school and home environments. We propose to proceed along two fronts: First, AIR and Educational Testing Service (ETS) will develop a new background questionnaire framework. Second, we suggest that the presentation of background questions be restructured. Specifically, we propose the use of a modular system, consisting of core and non-core modules. The majority of the core items will be used in all assessments, but some of the core items may be subject-specific and used each time a particular subject is assessed. These core items will be supplemented by non-core sets of items that may be used only once or that can be used periodically to track trends over time. These non-core items would be designed to collect data on variables of particular topical interest to educators, the public, and policymakers. Some of the non-core items may be subject-specific, and administered only to respondents taking a particular assessment (for example, mathematics-education questions).

We also discuss how we will improve existing items and phase in revised and improved items on the questionnaires, while still allowing comparisons with data from previous years. All survey items will be developed by experts, who have knowledge of how children and adults interpret and respond to survey items. All items will also be evaluated

extensively in AIR's cognitive laboratory, as well as through field tests and statistical analyses. The goal is to measure constructs as validly as possible, using the minimum number of questions, and, to provide data of a kind and in a manner that will be useful in informing debates about the relationship between specific educational practices and student achievement. In order to accomplish this, when choosing items for inclusion in the background questionnaires, we will always remain cognizant of the various reports that are to be produced.

The chapter begins with a conceptual discussion of the development plan and concludes with a description of the subtasks that will be completed under Task 11.

Overview of the AIR/ETS Plan for the Development of Background Items

New Model for Background Questionnaires

Current Model and Proposed Conceptual Changes

Currently, the process of developing background questionnaires begins with recommendations from the planning committee that constructs the framework for a particular subject-area assessment. As the assessment evolves, the development committee responsible for that subject area suggests additional background items that might be useful in providing a context for interpreting and reporting NAEP results. These recommendations, along with those of a background questionnaire advisory panel, are used to develop the four questionnaires currently administered along with each NAEP cognitive assessment: (1) a background questionnaire completed by every participating student; (2) a questionnaire completed by respondents' teachers at grades 4 and 8; (3) a school questionnaire filled out by the school principal or his or her representative; and (4) a questionnaire completed for each sampled student who has a disability or limited English proficiency. In the typical NAEP assessment, the total number of background questions across these four sources is fairly

large. The variables measured by these background questionnaires have been used for a wide variety of purposes. Some of the variables such as race/ethnicity, certain educational practice measures, and those thought to be related to socioeconomic status (SES) have been instrumental in producing useful documents (e.g., the instructional reports), but many of the variables appear only in summary data tables, where they receive little or no attention.

The Educational Statistics Services Institute (ESSI) recently reviewed the usage of NAEP items across 14 reports commissioned or written by the National Center for Education Statistics (NCES). Other than regular NAEP reports, the types of reports reviewed included *Research and Development* reports, focused reports on different subject areas, *NAEP at a Glance* reports, and *NAEP Facts*. The results of this review indicate that, despite the large amount of background information collected, many researchers examine only a small percentage of the variables available. Not surprisingly, gender, race/ethnicity, SES-related variables, and parents' education tended to be the most popular variables included in analyses of performance in any subject area.

The National Assessment Governing Board (NAGB), in its redesign document, has called for a reduction in the number of background variables collected. NAGB suggests that the student, teacher, and school background surveys would not necessarily all have to be collected as part of the same assessment. Instead, NAGB suggests the selective use of background questions and the development of reports that draw "on a core of background questions addressing the most essential issues" (NAGB, 1996).

AIR and ETS propose that the number of items in the background questionnaires be reduced by developing a small core module supplemented by a set of more flexible non-core modules, which may change from assessment to assessment. The core module would change only rarely, if at all.

The Current Population Survey As a Model

The Current Population Survey (CPS) collected by the Census and the Bureau of Labor Statistics provides a model of the sort we are suggesting. The CPS design combines the continuity needed for consistent reporting of key indicators over time with flexibility that positions CPS data at the center of many policy debates. Our proposal for the collection of background and contextual data for NAEP builds on the CPS model. The CPS uses a base survey, which is rarely changed, but is accompanied by supplements that address specific issues. Some CPS supplements, such as those on worker benefits and use of computers and technology, are asked on a regular basis, allowing policymakers to track trends over a period of time. Other supplements are designed to be used only once. CPS supplements, used extensively by researchers and policymakers, often provide the most up-to-date information available on social and economic topics.

The CPS model has many benefits. In the short run, the CPS is able to address issues of immediate concern to policymakers. For example, the CPS recently incorporated a series of items about job tenure, tobacco use, and food “security” (expenditures, sufficiency, etc.). As new issues arise, analysts refer to that database as a first source of insight. While the data may not be up-to-the-minute, many relationships are reasonably stable over time. For example, a supplement last used in 1994 gathered information about knowledge of the hazards of lead paint. An analyst interested in this issue can refer to the 1994 data, which may be the most current information available anywhere. Other benefits to using the CPS model are that in the redesign of the survey the developers have had to deal with changes in longitudinal items, an issue that will also be faced in NAEP (as discussed below). Additionally, cognitive laboratory work is used in CPS instrument development. NAEP may benefit from following this model.

Using Modules to Collect Background Data for NAEP

AIR and ETS propose that NAEP background questionnaires follow a model similar to the CPS. Each NAEP background questionnaire will

contain a short section of core-reporting variables, common to all subjects, as well as another short section of core-reporting variables that are subject-specific and will be administered only to students taking a particular assessment, or to their teachers and administrators. Additionally, the background questionnaires will have a relatively small number of well-selected, well-developed non-core items on carefully identified issues of current educational policy significance. In most cycles, at least some of these non-core items may be subject-specific.

The core module will consist primarily of the variables used for reporting initial NAEP results—variables such as gender, race/ethnicity, parents' education, and some school and community characteristics. Additionally, the core module may also include information on teacher characteristics, instructional practices, or school policies, either because there are variables that state policymakers may be particularly interested in monitoring over time, or because they may be critical to the proper specification of any other models that may be estimated with the NAEP data.

The non-core modules would allow for the collection of data on topics that are of current educational relevance. Some of the topics included in the non-core modules may be subject specific, such as questions relating to different methods of teaching reading. Other topics included in the non-core modules may focus on issues that do not relate to the content of a particular subject, such as the significance of class size. Some of the non-content specific modules may require modifying the NAEP sampling framework. For instance, a non-core background questionnaire designed to explore the educational context of charter schools would require oversampling these schools in order to obtain meaningful data linking student achievement with school characteristics.

Some of the non-core modules could be used at regular intervals to measure trends, while others may be used only one time.¹ These non-core

¹ Within a subject assessment in any given year, all participating students would take the same non-core modules. That is, while the non-core modules will vary year to year, they will not be matrix sampled within years.

NAEP modules would provide the same type of timely information on topics of interest to policymakers and researchers that the CPS provides, but with a focus on factors related to student achievement. In general, the choice of items for the non-core modules will be driven by decisions about report topics for that cycle, although NAEP might choose to collect trend data on a particular issue over several cycles, even though reports on that issue would only be prepared episodically.

Many important school-related issues could be examined in this way. For example, a fourth-grade reading assessment could gather in-depth data on reading instructional methods. Those data, linked to NAEP reading scores, could provide insight into the current debate over the use of phonics versus whole language or mixed reading instruction. Once a linkage between teaching approaches and student reading performance has been established, it may not be necessary to administer these items again, unless one were interested in tracking the prevalence of these types of practices over time. Other important educational policy issues that might be addressed in non-core modules include school organizational variables such as whether decisionmaking is centralized or decentralized, class size, school violence, bilingual education, and whether a district has adopted challenging content standards in the area under assessment. As major policy issues come to the forefront, new modules of questions could be developed, thereby providing maximum flexibility.

As is currently the case with NAEP, the contractor would need to stress the inability to draw causal inferences from such data. However, background data can provide context for NAEP results, and allow for more informed and useful speculations about patterns observed. Again, it is important to emphasize that we believe that a relatively few well-chosen, well-measured sets of variables related to important policy issues could contribute useful information about American education.

The change in the schedule of NAEP assessments from biennial to annual provides an excellent opportunity for implementing this system of rotating non-core background modules. Specifically, NAEP will have twice as many opportunities to place non-core modules before national samples;

this markedly increases the program's ability to track instructional and contextual trends.

Deciding on the Issues To Be Assessed

The education community and researchers from various disciplines have amassed an extensive amount of data on various factors that influence learning. In an ongoing project with NCES, AIR has conducted an extensive literature review on various factors of the home learning environment, in addition to SES, that are related to student performance. Figure 8.1 (pictured on the following pages) graphically displays a broad range of factors associated with home learning environment and shown to be related to educational achievement, along with examples of authors who have proposed, or who use, measures of these factors in their work.

The first notable feature of Figure 8.1 is its size. A substantial amount of research has been already completed. An organizing framework is needed to make sense of this amount of information. Figure 8.1 organizes this information hierarchically into three levels, with the highest level indicating general classes of associations with student learning (e.g., parent-child relationships, environmental stimulation). The second level characterizes various dimensions that have been posited as characterizing these classes of associations (e.g., extent of monitoring, types of parenting style, etc.). The lowest level lists the types of associations measured as indicators of these dimensions, along with examples of research that has used them.²

² Figure 8.1 shows that socioeconomic status (SES) is only one of the many factors listed. The other factors listed might be the mechanisms that account for differences in the achievements among students across different socioeconomic classes. When controlling for these other factors, the effect of associations with SES may disappear or decline. While these practices may be another set of class attributes (i.e. behavioral and/or attitudinal traits that vary with socioeconomic status), a survey instrument that provides information on these can help us to identify the specific aspects of SES affecting achievement.

The outline presented in Figure 8.1 provides only the start of a framework for analyzing factors that are important in student achievement. Importantly, the framework must be expanded to include factors in the school learning environment as well as those found in the home. As part of the preparatory work for developing NAEP background questionnaires, AIR and ETS will expand and complete this framework.

AIR is currently carrying out a project to explore structural models of the relationships of a wide range of school factors to school-level achievement. Part of this project involves linking the 1993-94 School and Staffing Survey (SASS) with 1994 non-NAEP state assessment files in 21 states and the 1994 state NAEP Grade 4 reading assessment. A school-level achievement measure has been constructed for 2,600 public elementary, middle, secondary, and combined schools in these states, making use of non-NAEP state assessment results to capture within-state variance and state NAEP results to capture between-state variance. A preliminary report on this study found that 35 to 40 percent of the variance in the school mean achievement measure was associated with a combination of a small number of factors, including student characteristics, teacher qualifications, class size, and school climate.

More sophisticated models, being implemented in the latter stages of the project, account for greater percentages of the achievement variance between schools. Among the more interesting findings is that the correlation between average class size and average achievement is highly significant, but only when the between-state variance component, based on NAEP, is included in the achievement measure. This result may be due to efforts to minimize variation in school resources and class sizes within their states.

The results of such analyses can contribute to the basis for selecting background measures to include in NAEP. For example, the SASS analyses indicate that a wide range of teacher opinions about what constitute serious problems in their school—such as tardiness and apathy to violence and poverty—are related to average achievement in the school. While these

correlations do not imply specific causal links, they provide evidence that must be considered in developing explanatory models relating school functioning and achievement.

NCES, possibly through consultations with NAGB, will make the decision about what policy issues would be the focus of each assessment.

Then, once a particular issue has been selected for study, the items to measure that issue will be developed by AIR in consultation with ETS and NCES. NCES will have final approval of all items. That is, an iterative development process is proposed, in which NCES identifies an information need (policy question), writers develop items and rationales, and NCES reviews and approves them. If NCES thinks it will be helpful, the writers could be present at the NCES deliberations.

NCES might also decide that some of the issues should be studied on a regular rotating basis, as is done with the CPS. Doing so would allow for the examination of changes in educational practices and help determine whether these changes are related to changes in NAEP scores for important subgroups or the total population.

Constructing Reliable, Valid, and Efficient Items

The modules containing the non-core items should be kept as short as possible, measuring only the constructs essential for research into questions that will be the topic of NAEP reports. Furthermore, each construct should be measured efficiently as well, perhaps with no more than five or six items. This small number of items can discriminate well along the entire range of the values of policy interest and do so with high reliability if carefully constructed. Efficient measurement requires that:

- constructs intended for measurement are clearly and appropriately tied to policy issues;
- each and every item elicits a valid measure of what it is intended to measure;
- the items elicit reliable responses; and
- when combined, the items cover the entire range for the constructs of interest.

Items that meet these criteria will yield data that will be useful to policymakers and interesting to the public and the media. Most of the NAEP background items do meet these criteria. However, there have been cases in which questions failed to reach such standards. A new series of reviews based on these specific criteria will be designed to ensure that each background question efficiently contributes to NAEP reporting. In addition, they will greatly increase the confidence with which we can draw inferences about the meanings of responses to background questions.

For example, there are sometimes inconsistencies between student and teacher responses on questions that the authors thought might have yielded concordant data. In the mathematics assessment, for example, student responses differed from those of their teachers in a number of cases, including the amount of homework assigned, frequency of mathematics worksheet use, amount of small-group work, frequency of report writing about mathematics or math projects, frequency of working with manipulatives, and how often students are asked in class to describe their problem-solving strategies. There may be some value in retaining this type of redundancy in questioning, since differences in perspective between students and teachers is an important area for research. That is, if these questions are yielding important information on the differences in how students and teachers view what goes on in the classroom, they may in fact be reliable, valid, and efficient items. However, if the difference in responses between students and teachers is caused by the fact that members of one or both groups are randomly responding, the questions should be rewritten or eliminated in future questionnaires. As discussed below, AIR

will use its Cognitive Survey Laboratory to assess these types of questions and revise them as appropriate. In addition, cognitive laboratories will give us confidence about the types of inferences that can be drawn from responses to operational survey questions.

The goal of the ETS/AIR team is to avoid problems by following the principles of efficient measurement noted above. In many cases existing background items already meet these criteria, but in some cases we will need to revise items to improve their measurement characteristics or write new items that measure constructs defined in the framework. All new and revised items will be evaluated using cognitive-laboratory techniques, a topic to which we now turn.

Using Cognitive-Laboratory Techniques to Improve Items

To understand how and why some survey items fail, fourth- and eighth-grade students and their parents or guardians have been brought into AIR's Cognitive Survey Laboratory. The students and parents responded to questionnaires that included items taken from NCES student surveys or from a proposed NAEP survey of parents. AIR staff trained the children to think aloud as they answered these items, so that they could understand the children's rationale for the responses they gave. Each child's parent or guardian was also asked to give the correct response for each question. When the child's answer differed from his or her parent's or guardian's, AIR staff asked follow-up questions to determine the reason(s) for the difference. Using this method, AIR found that some questions that appeared to be high in face validity actually were problematic.

Some questionnaire items failed because of comprehension problems, particularly among fourth-grade students. For example, when reading an item (not from a NAEP survey) about future educational plans that included an option about going to vocational school, some children mispronounced the word "vocational". AIR staff asked these children what they thought the word "vocational" meant. Many of the children defined the word as someplace you go to on vacations or schools where you go

during the summer. Clearly, any item that contains terms not understood by the respondent will be a source of error.

Some comprehension difficulties were caused by children's literal interpretations of items. When asked, "How often do you talk about things you have studied in school with an adult at home?," some children did not include conversations about school that took place during car rides with their parents. To these children, being in a car was different from being at home.

Some items were misinterpreted by children because of the grammatical construction of the items. For example, one item that used to appear on the NAEP student background questionnaire asks, "Does your father or stepfather live at home with you?." Children often interpreted this question as asking two different questions:

- Does my father live at home with me?
- Does my stepfather live at home with me?

If either question could be answered no, some fourth graders will answer the entire question negatively. After we divided the item into two separate questions, one about fathers and one about stepfathers, the error rate among fourth graders was negligible.

Another type of problem occurs when items do not measure what they are intended to measure because of retrieval problems. The simplest kind of retrieval problem is asking for some information that cannot be retrieved because the child has never learned the correct response. Since respondents try to be helpful, they will try to figure out the correct response. However, their logic sometimes does not match what is expected by item designers.

Many NAEP background questionnaire items ask about behavioral frequencies—such as how much television a child usually watches or how much time is spent on homework each school day. These types of questions

require judgment formation. The respondent must both retrieve information and synthesize it into a rate or frequency. Judgments of times can be influenced by many different factors. If something is not fun, children may overestimate the amount of time it took to do. In fact, in AIR's study in the Cognitive Survey Laboratory, about half the parents of eighth graders thought their children were overestimating the amount of time spent on homework each day.

Using the Cognitive Survey Laboratory, AIR staff also discovered that some children were making errors when responding to a few NAEP background questionnaire items because of problems in the response communication process. After respondents form a judgment, they must communicate that judgment to the survey administrators, usually by translating their judgment into a response option. Many survey items are closed-ended, providing respondents with a series of categories or choices. In some cases, these categories may overlap, or the category that best matches the respondent's judgment may be missing. In these cases, matching a judgment to a response choice is an imperfect process. Respondents may also alter their response because they think an alternative answer is more socially desirable.

Subtask 11.5.3 will describe in more detail how the background items for a specific content area will be reviewed using AIR's cognitive laboratory.

Avoiding Gender, Ethnic/Race, and Religious Bias in Background Items

The ETS/AIR team is experienced in drafting items that avoid gender, ethnic, class, and religious bias. All exercises will be subjected to a fairness review by ETS experts. The fairness review is described below in the introduction to Chapter 24.

Using Item Tryouts as Part of the Pilot Test

Pilot tests are essential in determining whether the items will function under real-world large-scale assessment conditions. They allow researchers to obtain numbers of responses that are infeasible in the cognitive laboratory so that quantitative analyses can be obtained. These range from simple analyses of item non-response rates and “don’t know” responses to analyses of item intercorrelations and response-distribution analyses. Where multiple items are used to measure a single construct, we can apply scaling techniques to select the most efficient set of items to measure each construct. For this, the items must elicit valid individual responses, and, combined, cover the entire range of the construct of interest.

Respondent debriefings should also be part of the pilot test process to allow for both the confirmation of the effectiveness of these items and the detection of problems associated with factors that could not be readily assessed in the cognitive laboratory. For example, administration in a group setting creates a different context than cognitive laboratories and so may influence item effectiveness.

The pilot tests will also be used for developing linking functions between the background items currently used for reporting purposes and revised versions of those questions. Since it is critically important that valid trend lines be reported, a strong statistical link must be established between the old and new reporting items. A strategy for developing the link and bridging to future versions of background questions is described under Subtask 11.4.

Specific Subtasks That Will Be Carried Out as Part of Task 11

The major deliverable for Task 11 is a set of modules of background questions. Because the modules will vary across assessments, Task 11 necessarily will run across the various assessment cycles. Furthermore, because the specific background items for a given assessment are to be chosen as part of Tasks 13, 65, and 123, there will also necessarily be a close relationship between these tasks and Task 11.

To understand the various subtasks described under Tasks 11 and 13 (see Chapter 10), it is useful to review the assessment cycle and the constraints that OMB clearances place on when work can be done. The field tests for the 2000 assessments must be carried out in January and February 1999.

As discussed in Chapter 10, Task 13, in anticipation of the 2000 background questionnaires, ETS has already begun systematically evaluating the items on the 1996 mathematics and science background questionnaires and has prepared revised and new background items for the 1999 field test. However, there will not be time to develop any additional questions for the 2000 background questionnaires after the new cooperative agreement is awarded. Therefore, the 2000 assessment will be transitional, incorporating questions from previous background questionnaires, some with revised wording that will be field tested in 1999. Additionally, a few new items will be field tested in 1999 and, if successful, incorporated in the 2000 background questionnaires. As part of the transition to a modular system of background questionnaires, we will identify some constructs measured in the 2000 assessment as core and others as non-core. Those that are identified as non-core will probably not be used in subsequent assessments, at least until some time has lapsed. The non-core items are likely, however, to figure prominently in the 2000 policy reports, which are discussed in Chapter 18, Task 49.

There will be less than one year from the time this cooperative agreement is signed until the OMB clearance package for the 2000 pilot test of new history and geography exercises is due. Although there will not be time to develop a complete set of non-core modules, we will have at least the fully-revised core module ready for pilot testing in 2000. Therefore, the 2000 pilot test and the 2001 assessment will be the first to fully incorporate the new background questionnaire design.

We now move to the discussion of the specific subtasks that will be completed as part of Task 11.

**Subtask 11.2. Develop Briefing Materials
for the Utilization Advisory Committee**

Working with NCES and NAGB, AIR will undertake a comprehensive literature review to identify important issues related to the subject area being assessed as well as important areas in the field of education more

generally. The literature review will identify topics that could be usefully addressed in the various modules. We will put a particular emphasis on identifying topics in which survey results can be meaningfully related to NAEP achievement scores.

Focus groups will be used to provide additional insight into topics of concern to stakeholders. Therefore, we will assemble groups of teachers, school administrators, subject-matter specialists, educational researchers, and parents to discuss topics that might be included in one or more of the modules for a given assessment area.

At this stage, we will not attempt to narrow the choices for possible topics for the core and non-core background modules, but instead will make the list as comprehensive as possible. One of the products of Task 13 (see Chapter 10) will be a framework to help conceptualize instructional practices, teacher background, school policies, and organizational structure and practice, frameworks similar to the one shown in Figure 8.1 for home background factors. We will also develop written materials evaluating the practicality of collecting useful information about each topic based on a limited number of self-reported questions (see Chapter 10).

All briefing materials will be given to NAGB and NCES for review,

All

briefing materials will be delivered to NAGB and NCES by March 1999.

Subtask 11.3. Develop Core Module

As discussed in more detail in Chapter 10, Task 13, ETS and AIR have already begun to review the current NAEP background questionnaire items under existing agreements. The current background questionnaire items will be pared down by eliminating redundancies and poorly performing items to create a transitional core background questionnaire for the 2000 NAEP assessment. This transitional questionnaire will be used as a starting point to develop a core background questionnaire for the 2001 NAEP that will also be used in subsequent NAEP assessments. Because of the modular structure of the background questionnaires, new items can easily be added to the non-core modules, making frequent changes in the core, non-rotating module unnecessary. Of course, the performance of the core items should be revised on a periodic basis to ensure that items continue to function as intended.

After reviewing the background items chosen for the 2000 NAEP “core” questionnaire, AIR and ETS will consult with groups such as the NAEP NETWORK and the NAEP subject-area committees to discover what constructs they think should be measured in the final version of the core module that will be developed in time for the 2001 assessment.

Once the constructs have been identified, we will examine the current set of measures and decide which should be rewritten or replaced and whether new items are needed. The decision about which items should be rewritten or replaced will be heavily influenced by the work AIR did with its cognitive review of NAEP background items for the Education Statistics Services Institute (ESSI), as well as the work ETS is currently doing in its review of the 1996 background questionnaire items. We will then review our suggested choices for the core module with NCES and NAGB. When approval has been given we will format the items for pilot testing in 2000 for use on the 2001 assessment. Preparation for pilot testing

entails obtaining OMB clearance in the summer of 1999 (see Chapter 22, Task 66).

The final version of the core module will be delivered to NCES by September 1, 1999, and will first be used in the 2001 national assessment.

Subtask 11.4. Maintaining a Trend as Survey Items Improve

As the core questionnaire is shortened and improved, the wording of some individual items will change to improve their reliability and validity. The wording of other individual questions undoubtedly will change in response to federal mandates. For example, OMB, starting with the 2000 census, will require that all federal surveys give respondents an opportunity to identify themselves as being from multiple racial or ethnic backgrounds. While this may provide better data, it will present challenges for maintaining trends as the classification variables change. In general, we propose a phase-in that will use the NAEP pilot test to try out changes to survey items, and will design statistical linking models and methods of estimation to ensure maintenance of trend lines.

AIR has been working with the application of IRT-like models to the problem of maintaining trends in a target variable when classification variables change. We propose to work with NCES to continue to refine these models and adapt them to NAEP, where the target variable is only partially observed. Relying on data from the field test conducted in 1999, we will refine and adapt the statistical models. We will explore the precision of the estimates and evaluate the requirements for a reliable linkage, such as sample size and statistical constraints.

If, as we expect, we find that a reliable linkage requires a large sample size, we will recommend including both old and new versions of the items on operational NAEP when new items are being considered for inclusion on the subsequent administration. For example, if we plan to use the new race/ethnicity item in 2001, we may recommend treating the operational NAEP in 2000 as a “linkage sample.” The use of such bridging models will

be conducted in a manner that does not slow NAEP reporting. Those analyses will not be placed in the critical path in any year but will be used as bridges to future assessments.

We envision the linkage process as ongoing, since both core and non-core modules may require time trends. Below, we provide a brief outline of our general approach. A more detailed and technical discussion of the model and its extensions is presented in Appendix A.

Consider a survey item, x_1 , that has its wording changed from one assessment to the next to yield variable x_2 . Also, suppose that we are trying to maintain a trend in which this item is used as a classification variable, and we are interested in keeping track of the subgroup averages of a target variable y . After some linkage study, we may want to project the newer version of the item backward to express the trend in terms of the new variable x_2 . This provides a trend line that could be continued into the future using the new item.

The essence of the task is to describe the conditional distribution $f(y | x_2)$, even though we only observe $f(y | x_1)$. The task requires an overlap sample and some very strong assumptions. Several types of linking samples may be envisioned, for example:

- measure x_1 and x_2 for the same sample of individuals at a fixed point in time;
- measure x_1 and x_2 at the same time in randomly equivalent groups; or
- measure x_1 and x_2 for a single sample of individuals, but at different points in time.

Regardless of which one of these three alternatives is used, by treating both old and new variables x_1 and x_2 as measuring the same underlying trait without bias, but recognizing that both include some measurement error, we can use a classic two-parameter IRT model with a normal rather than a logistic distribution function to estimate not only a

location and discrimination parameter for each item, but also a regression parameter (or parameters) of the regression of the latent trait on the survey's target variable. In this example, we assume that the measurement error is independent of y , and that the trait of interest accounts for all of the variance that the two items share. This model can be estimated using one of the three linking samples mentioned above.

In projecting the relationship between y and x_2 back to surveys that collected x_1 , the sample model attributes changes in the observed relationship to changes in the relationship between the latent trait and y . Thus, using the target data set, the relationship between the latent trait and x_1 is held constant, and new values for the regression parameters are estimated. In the target data set, however, the scale and location of the latent trait are no longer arbitrary—they are determined by the parameters describing the relationship between the latent trait and x_1 . Hence, in addition to a new regression parameter, we can estimate a constant term. Thus, shifts in the latent trait or its relationship to y induce changes in the inferred relationship between x_2 and y .

We can expand this basic model in several important ways. First, rather than a measured target variable, we will have to extend the model to partially observed target variables and, ultimately, partially observed multivariate target variables.

Second, it is probably possible to incorporate additional information into the link functions about item placement to test for context effects. Also, the current model assumes that the latent trait incorporates the entire relationship between the individual items and the target variable. This may not be the case: In addition to the shared relationship to the target variables, it is possible that the “measurement error” is also related to the target variable. It may be possible to establish models that estimate and test for such relationships.

It is also likely possible to adapt this method to linkages based on randomly equivalent groups. Further theoretical development would be

necessary to identify the assumptions that would be required, and empirical analysis would be required to evaluate the plausibility of those assumptions.

Finally, models of the linking error will be required. These are necessary to identify the sample sizes required for linkages, and for estimating the standard errors associated with statistics inferred through the linkage.

Appendix A (under Tab 57) provides further discussion of the theoretical basis of this model.

Subtask 11.5. Develop Modules of Non-Core Background Items for the 2001 and 2002 Assessments

After the UAC decides which topic or educational issue to include in the 2001 and 2002 assessments, the actual preparation of survey items to be included in the module will occur in three steps:

- review existing items to identify those to be retained in their existing forms;
- draft new items where constructs need to be measured and no items are available; and
- refine the items through cognitive testing.

The text below describes each of these phases.

Subtask 11.5.1: Review Existing NAEP Items

As described in Tasks 13, 65, and 123, AIR will identify items from previous NAEP background questionnaires that measure constructs

NCES identified for inclusion in the 2001 or 2002 non-core modules. When possible, we will incorporate items from previous NAEP questionnaires that are not being used in the core background questionnaires into the non-core modules, to allow researchers to measure trends over time. Items from previous NAEP questionnaires that are being

considered for inclusion in a non-core background module will be evaluated for coverage, quality, and efficiency.

This process of evaluating items from previous background questionnaires has already begun under existing agreements. As discussed in Task 13 (see Chapter 10) below, the 1999 field test will provide data that will allow us to evaluate newly revised questions. We will continue that work under this task, with the goal of completing the evaluation by spring 1999 for inclusion of some items in the 2000 pilot test and others in the 2001 pilot test. Members of the NAEP NETWORK and of the subject-area committees will be asked to review all questions prior to their inclusion in final modules. NCES will have final sign-off authority.

As discussed in Subtask 11.5.3, before pilot testing new items, AIR will use its cognitive laboratory to test all items from the existing NAEP that may be used in future non-core modules to ensure that they efficiently and accurately measure the constructs that they are intended to measure. Items that fail to meet these criteria will be reworded and retested.

Subtask 11.5.2: Draft New Items for Non-Core Background Modules

The AIR staff has expertise in drafting items for a variety of surveys, including surveys in which the respondents are children of various ages. These experts will draft new items for inclusion in the non-core modules. AIR's experts are knowledgeable about the craft of questionnaire development and possess an understanding of the questionnaire response process, the people who will be answering the items, and how these items will be used.

Item design is an iterative process and expert review is a critical part of the process. AIR staff will appraise each new item based on their knowledge of the response process. They will use an item appraisal form that compels consideration and evaluation of each item with respect to dozens of factors known to adversely affect comprehension.

Item drafting for non-core items is an ongoing process that will be done in response to suggestions by NCES. Items for at least one non-core module will be drafted in time to be evaluated in the Cognitive Survey Laboratory and included in the 2000 pilot test and in the 2001 operational assessment.

Subtask 11.5.3: Cognitive Testing of New and Existing Items Using the AIR Cognitive Laboratory

After existing items have been selected and new items have been developed, AIR will refine the items using its Cognitive Survey Laboratory, which is based on a model of the questionnaire response process developed by Sudman, Bradburn, and Schwarz. Briefly, this model specifies that, in order to answer a survey question, the following cognitive activities must be performed:

- comprehension/interpretation of item,
- retrieval of item information,
- formation of judgment, and
- edit of answer.

To investigate the questionnaire response process, AIR uses think-aloud protocols, directed probing, paraphrasing, and projective techniques. These techniques provide an understanding of the reasons for respondent errors. This knowledge, linked to an underlying model of the response process, will enable AIR to make recommendations for modification of specific survey items to permit the collection of valid data.

The Cognitive Survey Laboratory is an effective device for rapidly and inexpensively identifying problem items, as well as identifying the reasons why items are not working as intended. Certain tools such as “think alouds,” comprehension assessments (using probes, projective, and paraphrasing techniques), and thorough debriefings are critical elements of this work.

AIR will also use the Cognitive Survey Laboratory to systematically analyze all new and existing questions from the NAEP background questionnaires. The goal will be to identify the items that best measure their intended constructs.

Subtask 11.5.4: Prepare for Pilot Test and Evaluate Pilot-Test Data

Our general goal is to avoid introducing new items in the operational NAEP until their validity and reliability has been proven. This is particularly important when dealing with items for which we are interested in maintaining a trend. However, the 1999 field test for the 2000 assessment is covered under a previous cooperative agreement and the field-test items will be finalized before the new cooperative agreement takes effect. As a result, we anticipate that the first use of the non-core modules will occur in the 2001 assessment. Even then, we will not have time to create an entire set of non-core modules by the 2000 pilot test, but we will have readied the set that we intend to use in the 2001 assessment.

Preparation will entail obtaining OMB clearance (see Chapter 10, Chapter 22, and Chapter 37) and printing the booklets.

After the pilot test, we will evaluate the completeness of the data as well as the precision and efficiency of the items employed. The particular analyses employed will depend upon the items being evaluated. Approaches are likely to include:

- analysis of nonresponse and skip patterns;
- consistency checks among multiple respondents (e.g., students in the same class, students and teachers); and
- analysis of the efficiency of multiple measures when they are used.

Initial analyses will examine the extent of item nonresponse. These analyses will identify the items that respondents are unwilling or unable to answer. Ideally, such items will have been identified and modified or

deleted during the cognitive testing. However, the larger pilot-test samples and a group mode of administration may identify some previously undetected problems. We will review any such items to identify potential sources of nonresponse and work with NCES to either revise or omit them.

At times it is possible to assess the consistency of responses across respondents. For example, items asked of students and teachers could be used to confirm one another. When appropriate, we may include items on the field test forms explicitly for confirmatory purposes; these items would not be used on operational forms. For example, if a piece of information can be obtained as reliably from one teacher as from five students, efficiency suggests that the operational item be asked only of the teacher.

The reliability of survey items can also be examined when multiple indicators are collected to measure a single trait. IRT models are useful in this regard because, in contrast to traditional reliability measures, they provide estimates of the reliability of a scale at different levels. For example, Cohen, Bohrnstedt, and Jiang (1996) found that many home background factors measured in NELS:88 were only measured well within very narrow regions. We will use these methods to construct final surveys that use as few items as possible to yield valid, reliable estimates along the range of interest on the scales.

A set of final survey instruments for the 2001 assessment will be prepared for OMB clearance by May 1, 2000. These will be accompanied by a complete report on field testing activities and a field test evaluation report for the 2001 assessment.

Subtask 11.5.5: Pilot Testing Methods of Linking Different Versions of Questions

AIR staff will pilot test methods used to maintain trends when the items measuring key background data are changed. For example, as noted above, government-wide guidelines now dictate a different format for measuring race than has been used in the past. In addition, the ESSI “home

background” task has proposed substantive changes in the ways that several basic items are asked of fourth graders. Since some of these items are used in reporting trends, it is important that methods of maintaining trends are established and proven before they are installed on NAEP’s critical path. These linking methodologies are described in detail in Appendix A (under Tab 57).

CHAPTER 9.

CONDUCT SEMINARS ON PSYCHOMETRICS

Task 12

EXECUTIVE SUMMARY

Objective: To provide a series of forums at which stakeholders can learn about NAEP methodologies and discuss NAEP findings. To meet this objective, Educational Testing Service (ETS) and American Institutes for Research (AIR) will:

- meet with the National Center for Education Statistics (NCES) to plan and implement a series of seminars on key NAEP topics, and
- make logistical arrangements for all seminars.

TASK 12.

CONDUCT SEMINARS ON PSYCHOMETRICS

Introduction

As NAEP continues to increase in complexity, scope, and importance, the demands on those charged with managing the project increase as well. If the program is to be managed effectively, it is important that all participants are cognizant of the key operational, measurement, exercise development, and reporting issues surrounding large-scale assessments such as NAEP. Educational Testing Service (ETS) and American Institutes for Research (AIR) welcome the opportunity to assist the National Center for Education Statistics (NCES) in convening seminars and training sessions for NCES, National Assessment Governing Board (NAGB), and contractor staff.

While it is our understanding that the seminars and training sessions will focus primarily on topics of interest to these groups, ETS and AIR also anticipate that there may be public meetings attached to the seminars which policymakers, education officials, the general public, and interested organizations will attend. We encourage this wider participation. Because NAEP is a national program with policy and educational implications for millions of American educators and students, ETS and AIR feel an obligation to keep as many audiences as possible informed about NAEP activities. The public portion of the seminars will contribute to a broader understanding of what NAEP is, how it operates, and the ways in which it can illuminate public policy debates. Additionally, public seminars will contribute further perspectives to the project from both presenters and participants.

Seminar/Training Session Topics

Frequently, project staff are invited to make presentations about NAEP to many groups across the country. ETS and AIR staff members also regularly publish articles in scholarly journals. The two organizations will use this experience in planning the seminars and training sessions. It is our expectation that the selection of the final set of topics will result from a collaborative interchange among all interested parties. Staff members will commit their best efforts to the seminar planning, so that the most timely topics are included.

NAEP has an extensive database and consistently explores vanguard technical and theoretical approaches to exercise development, assessment, and analysis. These explorations can provide valuable information to seminar participants. Among the seminar topics that we suggest for consideration are:

- what NAEP has learned about conducting performance assessments,
- tailoring NAEP reports to diverse audiences,
- linking other assessments to NAEP,
- scoring NAEP assessments,
- holistic, analytic, and primary trait models of constructed-response scoring,
- using NAEP findings for teacher training and curriculum reform,
- statistical methods for measuring item bias,
- statistical methods for controlling rater effects,
- developing NAEP exercises,
- the NAEP long-term trend assessments,
- using NAEP data for international comparisons,

- computer-based testing and NAEP,
- NAEP and the Worldwide Web,
- "market-basket" reporting approaches,
- reconciling NAEP's diverse goals: measuring what students can do versus providing a stimulus to curricular reform,
- using item response theory models in NAEP analysis,
- the use of plausible values in NAEP analysis,
- linking NAEP data to that found in other surveys,
- results of NAEP cognitive laboratory studies,
- different approaches to generating NAEP proficiency estimates,
- making NAEP reports more understandable: new modes of data presentation, and
- issues surrounding the inclusion of scores achieved under non-standard conditions (e.g., accommodations for students classified as IEP and LEP) in the NAEP proficiency estimates.

This list of topics is not meant to be exhaustive, and a final list of topics will be set after consultation with NCES.

Seminar/Training Session Formats

There is a tremendous potential for creating interactive meeting formats in which the audience can participate and share questions and knowledge. ETS and AIR will work with the NCES staff to develop meeting activities that will stimulate discussion and the continuing exchange of ideas. In addition, ETS and AIR will work together to produce written seminar materials which will be delivered to participants well before the date of the session.

ETS and AIR propose to conduct five seminars, approximately one every year over the life of this cooperative agreement. We further suggest that outside assessment experts be commissioned to participate in two or three of the sessions. These experts would be asked to make presentations on topics of interest to NAEP participants. A proposed list of participants will be submitted to NCES with the plan for the seminars. Selection of the participants and the final seminar topics will be completed mutually by ETS, AIR, NCES, and other organizations that NCES wishes to have included.

The reviewer should note that some funds for seminars and training are available under the current NAEP cooperative agreement. Rather than assume that we should effectively double funds for seminars, we will assume that new funding for seminars and training needs to commence only at the conclusion of the existing agreement.

Meeting Arrangements

We anticipate that all meetings will be held in Washington, DC or the surrounding suburbs.

The proposed collaborative effort to prepare sessions that are productive and useful for all participants is a high priority for ETS and its partners, and an important component in sharing NAEP program information.

CHAPTER 10.

FINALIZE THE 2000 NAEP BACKGROUND QUESTIONNAIRES AND PREPARE IMT/OMB CLEARANCE PACKAGE

Tasks 13 and 14

Objective: To select and finalize background questionnaire modules for use in the 2000 NAEP assessment. To achieve this goal, American Institutes for Research (AIR) and Educational Testing Service (ETS) will:

- continue a systematic review of background questionnaires begun for NAEP in 1998,
- select final variables for 2000 after reviewing question performance on field tests and past assessments, and setting plans for 2000 policy reports, and
- prepare and submit all materials needed to obtain clearance for instruments from the Department of Education's Information Management Team and from the Office of Management and Budget.

INTRODUCTION

As discussed in Chapter 8 (Task 11), the requirement for Office of Management and Budget (OMB) clearance means it will not be possible to develop new items for the 2000 assessment under this cooperative agreement. Instead, the Current Population Survey-like model for the background items will begin with the 2001 assessment (see Chapter 22, Tasks 65 and 66). However, the number of background items can and will be reduced as part of Task 13. Furthermore, for the 2000 assessment, the items chosen will be shown to fit into a new and more elaborated framework than has guided background question development in the past.

The three subject areas that will be assessed in 2000—mathematics, science, and reading—have all been assessed using the current frameworks. Consequently, sets of background questions exist for each assessment. Furthermore, work has been ongoing by Educational Testing Service (ETS), under the current cooperative agreement, and by American Institutes for Research (AIR), under separate task agreements with the National Center for Education Statistics (NCES), to evaluate the validity, reliability, and utility of the current questionnaire items. We propose to continue this work, with the goal of achieving an efficient set of measures for constructs related to the teaching and learning of mathematics, science, and reading.

As described in Chapter 8, which presents an overview of our general approach to the development of the background questionnaires, it is our position that new or revised background items should always be pilot tested before operational use to confirm expected item functioning. Consequently, any substantial efforts to revise or replace background items must be initiated well in advance of pilot testing, to allow timely development of pilot-test clearance packages. For the 2000 assessment, this longer-range

work is occurring under a previously existing cooperative agreement.

ETS will use the 1999 field test to evaluate revised background questions. As discussed in more detail in Task 13 below, new versions of some of the background questions from the student, teacher, and school questionnaires will be tested. In addition to testing reworded questions from previous questionnaires, a small number of new background questions may be field tested as well. AIR and ETS will use the results of the 1999 field test to help finalize the selection of background items for the 2000 assessment. In particular, we will shorten the background questionnaires by eliminating redundancies and poorly performing items.

Early specification of the policy reports that will be written for the 2000 assessment, as discussed in Chapter 18 (Task 49), is important for the work of background questionnaire development. Beyond the variables used to define the reporting categories required by law, the most important constructs are those that will be utilized in the reports. Therefore, as discussed in Chapter 8, the development of report outlines and background variables for any given assessment cycle ought to proceed in tandem.

Finally, AIR and ETS will hold a full series of discussions with users of NAEP data and reports. The goal of these discussions will be to identify a parsimonious set of educational policy and practice variables for which stakeholders are most concerned with tracking trends from previous assessments.

As discussed above, the process of compiling information and drafting a set of recommendations for the 2000 background questionnaire has already been started by ETS under the current NAEP agreement in preparation for the 1999 field test. We will consult closely with NCES as we continue this process and develop

the final set of questions for the 2000 background questionnaires.

Once these deliberations are concluded, we will proceed to the last step, specified in Task 14, the development of the clearance package for review by the Department of Education's Information Management Team (IMT) and the Office of Management and Budget (OMB).

TASK 13.

FINALIZE THE 2000 NAEP BACKGROUND QUESTIONNAIRES

Development of final questionnaires for use in 2000 will proceed in stages, described below: 1) classify items into a systematic framework and evaluate analytically, 2) use reporting plans and input from key stakeholders to drive the final selection of background items,

Classify Items into a Systematic Framework and Evaluate Analytically

The most recent operational administrations of the mathematics and science assessments in 1996 involved 22 distinct, but overlapping, background questionnaires and hundreds of individual items. The instruments included 10 student questionnaires differentiated by grade and subject area and three school questionnaires differentiated by grade. They also included mathematics and science teacher questionnaires for fourth-grade and eighth-grade teachers, a questionnaire for teachers of students in the twelfth-grade advanced mathematics sample, and a questionnaire to be completed for each student with a disability or classified as limited English proficient, whether included or excluded from the assessment. The situation for reading, last administered in 1998, is similar.

Each questionnaire addresses a wide range of constructs related to teaching and learning and, in some cases, constructs are measured redundantly across, as well as within, instruments. Most notably, the student and teacher questionnaires contain a number of parallel questions about the specific instructional practices to which the students are exposed. In addition, a variety of SES

measures are spread across the student and school questionnaires, and overlapping questions about extent of subject-area instruction appear in both the teacher and school questionnaires.

More recently, a number of the student and teacher background questions have been evaluated based on observed response frequencies and subjected to cognitive analysis using think-aloud protocols and, wherever possible, external validation by another data source. A limited set of item modifications, based on these analyses, were incorporated into the 1998 assessment.

ETS has systematically reviewed all the questions on the 1996 mathematics and science background questionnaires in preparation for the 1999 field test, and more changes in background item wording are being evaluated. The questions in the 1996 student, teacher, and school questionnaires were reviewed by three members (for each subject) of the Math/Science Standing Committee, who contributed written comments. In addition, several individual reviewers evaluated the background questions. All the reviewers' comments were compiled and discussed by the three-member committees in a series of conference calls. The recommendations from each of these committees then went to the appropriate instrument development committees in early May. The instrument development committees responded to the recommendations from the reviewers and sometimes suggested additional items or deletions. ETS has begun a similar process to evaluate background questions from the 1998 student, teacher, and school administrator reading background questionnaires.

These revised items, and new items as well, will be evaluated in the 1999 field test. In some cases two different versions of a background question will be tested, to see which version produces the most efficacious and valid range of responses. ETS will limit the 1999 field test to evaluating revised and new questions, since

enough data already exist to evaluate questions whose wording is not being changed.

AIR and ETS will use the results of the 1999 field tests to draft the new 2000 background questionnaires specified under this task. Using a system similar to that illustrated in Figure 8.1 (shown in Chapter 8), we will develop a comprehensive and research-based framework which will be used to classify all of the existing background items as well as the new variants developed for the 1999 field test. A subset of these items will then be incorporated into the 2000 background questionnaire.

As items are sorted into constructs, the unevenness of coverage will be readily apparent. Unsurprisingly, many possible constructs in the framework will not be measured at all, since some constructs do not lend themselves to effective measurement in large-scale surveys. On the other hand, some constructs will be covered by a large number of discrete questions. For example, the grade 4 School Characteristics and Policies Questionnaire used in 1996 included at least 29 separate questions related to school climate, including parental involvement, school disruption, and the mobility of the student population.

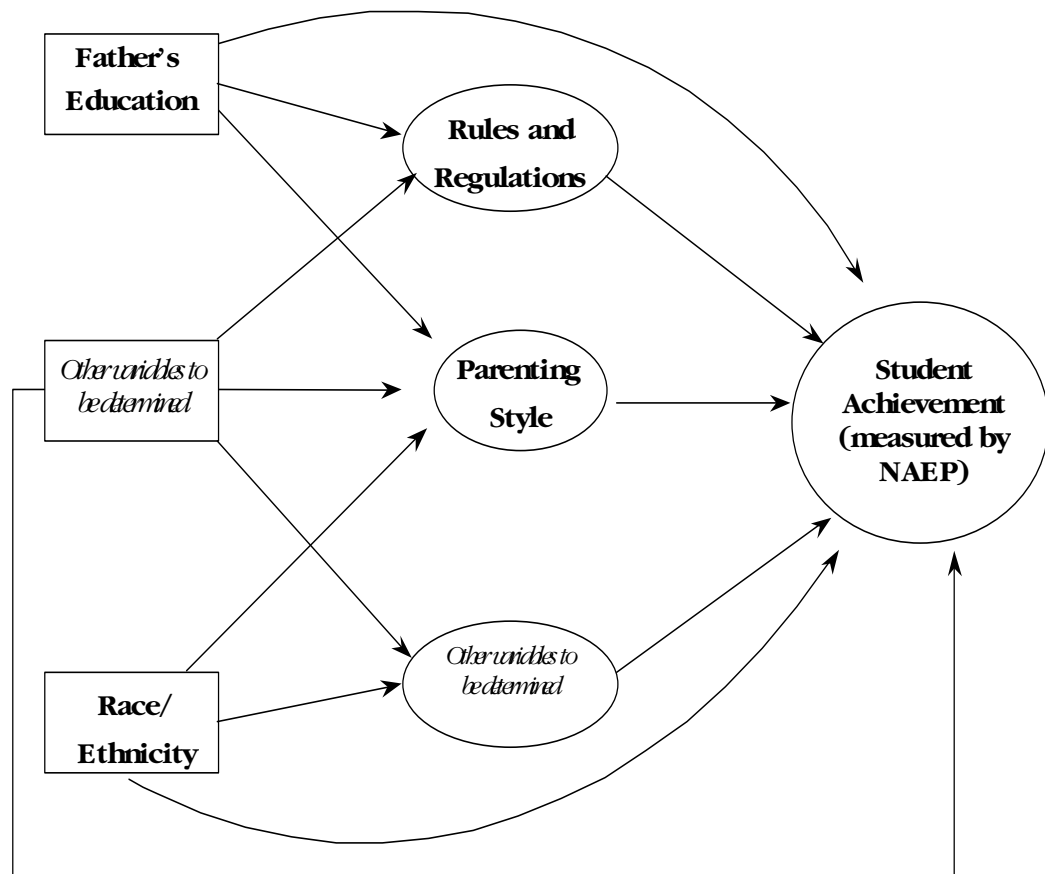
Given the relatively late stage at which this cooperative agreement takes effect relative to the development of the 2000 background items, our goals for this particular assessment cycle are limited. AIR and ETS will focus our review on paring the number of measures of the same construct. As already noted, we will not attempt to replace or revise items unless the new variants have already been included in the 1999 field test. Moreover, we do not expect, in most instances, to excise entire constructs.

A number of sources of information will be used in identifying candidates for deletion. Key among them will be analyses of how well the questions “fit to the model,” using data

from previous administrations and/or the 1999 field test. The new framework will allow us to hypothesize relationships between measures of the same construct, between constructs, and between each construct and student achievement. These relationships will then be evaluated empirically, and, where there is redundancy, only one set of measures will be proposed for retention.

An example of the type of approach to be taken in evaluating items for inclusion can be seen in Figure 10.1. One begins with a measurement model in which one determines how well individual items relate to the construct putatively measured. After confirming that items fit important framework categories, they are entered into a structural equation model, the relationships of the constructs to achievement are estimated, and the overall fit of the model evaluated. The variables in Figure 10.1 are, of course, included only as examples.

Figure 10.1
Structural Model
Effect of Background Variables
on Student Achievement



In a separate analysis, we will seek to identify constructs that are measured by current NAEP background items, but have not been used in any published reports or articles during the past decade. Both official NAEP publications and reports or articles authored by secondary data users will be considered in this analysis, and any constructs that have been neglected during this period will be considered for deletion in 2000. The orphaned constructs, if any, may have been passed over because of the limitations of the measures used to gather data, rather than because of an inherent lack of interest in the topic area. In that case, the constructs are potential candidates for reintroduction at some subsequent assessment cycle, probably in a non-core module, when there are better opportunities for the development of new measures.

Use Reporting Plans and Input From Key Stakeholders to Drive the Final Selection of Background Items for 2000

The analyses described above are intended to derive the best available measures for a given set of constructs (given the pool of available measures) and to eliminate a few of the most egregiously underutilized constructs. Two other activities can play a more positive role in shaping the final set of 2000 background questions: early planning of 2000 reports and consultations with stakeholders. As described in Chapter 8 (Task 11), the efficiency and effectiveness of NAEP can be substantially enhanced by an integrated system of instrument development and report design.

The advantages of this approach will be even more apparent in subsequent assessment cycles, when the issues-based modules are fully deployed, and when it will be possible to construct new items in support of the proposed reports as well as to simply delete unneeded items. However, even in the 2000 assessment cycle, which constitutes somewhat of a transition between old and new models of development, it will be advantageous to reach agreement

with NCES as early as possible on the outlines of the reports, particularly the policy reports. The process for developing these reports is elaborated in Chapters 18, 34, and 47.

In addition to more targeted reports, we will incorporate a number of innovations that will make data more readily accessible in a variety of formats. In particular, new table-generation and analysis software will allow interested users to conduct specialized analyses or to produce their own tailored reports (see Chapters 7 and 19, and the *NAEP Interactive* section of the compact disk sent as a companion to this proposal). We also propose that the complete text of all background questions be available on the Web, and that links be developed between those questions and relevant parts of the summary data tables.

Assessment and curriculum specialists in state departments of education seem among the most likely consumers of these reports and services. Consequently, it would seem prudent to consult their interests before defining a final set of background questions for the 2000 assessment. If specific constructs are of particular concern to the states, then these constructs also should be strong candidates for inclusion. For example, if states are interested in monitoring their own progress in meeting the NCTM standards, we would want to include questions about these standards on the teacher or school administrator questionnaires. Of course, as with all interest groups, our concern must also be with helping state participants identify a limited set of top priority choices rather than allowing the set of desired constructs to expand indefinitely. To establish a dialogue with the states on this topic, we will work with NCES to make use of the forums provided by the NETWORK and EIAC.

TASK 14.

PREPARE IMT/OMB CLEARANCE PACKAGE

AIR and ETS will prepare all materials needed for the formal clearance and approval process for the final student, teacher, and school questionnaires according to the requirements of NCES, IMT, OMB, and the National Assessment Governing Board (NAGB). Based on past experience, we expect that the background questions will not have been dramatically revised after the 1999 field test. Therefore, only changes to the background questionnaires made as a result of the field test will be submitted to IMT and OMB for the 2000 assessment.

ETS and AIR recognize that the Application for Cooperative Agreement indicates that the forms clearance package for the 2000 assessment should be submitted to NCES not less than four months prior to the start of the printing of the background questionnaires. Hence, we propose the following series of timelines and steps for clearance of the instruments. The dates in this schedule represent final due dates. In some cases, materials may be submitted earlier.

- April 21, 1999 – ETS/AIR will submit to NCES for review and approval copies of all background questionnaires and data-collection forms, as well as draft materials and forms required by IMT and OMB for clearance. ETS and AIR staff will be available, if requested, for presentations to NCES, NAGB, NCES consultants, and the NCES Interdivisional Review Panel.
- May 1, 1999 – ETS/AIR will make revisions to questionnaires and other materials and provide NCES with 10 copies of Volumes I and II¹ of the forms clearance package for the IMT/OMB and NCES/NAGB reviews. The

¹ Volume III of the clearance package, which will include background items to be pilot tested in association with new exercise pilot testing in history and geography, will be submitted by June 1, 1999. The development of this volume is described in Chapter 22 (see Task 65).

content of each volume of the clearance package is described below.

- June 1 – July 31, 1999 – ETS will submit the cognitive item clearance packages to NCES and NAGB for review. The production of these volumes is covered under an existing cooperative agreement.
- No later than September 1, 1999, ETS will receive final OMB approval and begin printing.

Throughout the entire clearance process, ETS and AIR staff will be available, if necessary, to join the NCES project officer for a review of the materials with the NAGB, OMB, and IMT staffs.

Much of the success in obtaining OMB clearance depends upon the quality of the submitted clearance package. The package must be brief, yet complete, and explicit as to research design and respondent burden. The package must contain a complete explanation and defense of the proposed sampling plan, and data collection and analysis techniques. Other principles for developing a successful OMB package are to write clearly, to avoid technical jargon to the greatest extent possible, and to provide sufficient context that the package is comprehensible even to reviewers unfamiliar with the project.

The clearance package must include a completed copy of OMB Standard Form 83-I (*Instructions for Requesting OMB Approval*)

under the Federal Reports Act, as Amended), including the certification statement signed by an OST official; a copy of each data collection instrument; a brief description of each instrument in general, and for each related group of questions contained in the instruments; a brief description of the purpose of the data being collected; and a supporting statement.

The information in the clearance package for the 2000 assessment will include item-by-item justification for all the background questions that were changed as a result of the 1999 field test, as well as justification for each item by itself and as part of a major conceptual area. It is important to note that items will be included in the background questionnaires because they relate to a specific conceptual area as discussed both in this chapter and in Chapter 8.

ETS and AIR will submit the background questionnaires and data collection forms for clearance in two separate volumes for ease of review. Subject-area cognitive questions will be submitted separately. (For the 2000 assessment, the subject-area work is being performed under a separate cooperative agreement.) The two volumes discussed in this chapter are:

Volume 1: Supporting statement describing the reasons for the study and containing forms and information required by the Office of Management and Budget for the clearance process, including the following:

- a statement supporting the necessity for the data collection,
- a description of the National Assessment of Educational Progress and the purpose of the information being collected,
- a discussion of efforts to identify duplication,
- a discussion of how respondent burden is being minimized,

- a discussion of the confidentiality of responses,
- special justification for any “sensitive” items (note that detailed justification of all items to be included in the survey forms will be included in Volume II),
- a table showing the data collection form, estimates of the response and preparation time, total number of respondents, and total burden (table notes will indicate the formulas used to calculate burden),
- a discussion of the estimated federal and respondent cost of the data collection (and how these were calculated),
- a schedule showing the timing of each data collection and reporting of analyzed results,
- a description of the statistical methods that will be used in the collection of information including:
 - sample design specifications (a description of the potential universe of respondents and sample size, and any stratification procedures used for identifying potential respondents)
 - a discussion of methods for handling non-respondents to maximize response rate
 - a brief summary of the planned analyses
 - the steps taken to have statistical aspects of the study design reviewed by non-agency personnel,
- other information required by NCES, IMT, OMB, and NAGB, and
- cover letters and information packets.

Volume II: Conceptual Framework and Background Questionnaires includes the following:

- the conceptual framework, developed by the UAC and ETS and AIR staff members, which describes the areas that will be addressed by the 2000 assessment and provides a

rationale for including the specific items in the background questionnaires, and

- copies of the student, teacher, and school questionnaires; IEP/LEP student questionnaire; and data-collection forms.

ETS and AIR will prepare and assemble the above materials in a clearance package and will highlight changes from the pilot-test clearance package. Since each package is reviewed by a number of individuals—with each person often looking for a specific discussion—the layout must enable reviewers to access the information easily. Tables of contents are an essential component of our packages and the different questionnaires will be color coded to facilitate differentiation.

Volume II will be submitted in a three-ring binder to allow quick replacement of items. All changes requested by NCES, including those generated in OMB or IMT reviews, will be made by ETS and AIR and submitted electronically to NCES. Both Volume I and II will be submitted no later than May 1, 1999.

In summary, we will take all steps necessary to obtain clearance on schedule for NAEP. The development of high-quality surveys and well-organized clearance materials are central to the success of these efforts.

CHAPTER 11.

PREPARE FOR MATHEMATICS MARKET BASKET

Task 15

Objective: To develop a market-basket form in mathematics at grade four, and to report results based on that market basket in conjunction with the 2000 assessment. To meet this objective, Educational Testing Service (ETS) will:

- design a market basket that will represent the types of exercises included in the NAEP mathematics assessment,
- pilot test exercises for this form in 1999,
- build two operational market-basket forms for 2000, one composed of newly developed exercises intended for release and the other of existing NAEP exercises,
- administer these forms in conjunction with the 2000 assessment, and
- analyze the results of the study and write two reports, a popular report and a research and development report.

INTRODUCTION

Originally proposed by Educational Testing Service (ETS) short forms or market baskets designed to complement the main NAEP assessments offer many possible advantages. As discussed below, market baskets and short forms are distinct concepts. The former are reporting mechanisms, while the latter are instruments to be administered to students.¹ In assessments like NAEP both short forms and market baskets may have uses, and their roles may interact. Market baskets—which would contain examples of the types of exercises and content included in the main NAEP assessment—could serve a number of purposes. First and foremost, the market basket would provide a valuable new way of giving meaning to student achievement on NAEP instruments: Average student performance on the overall assessment and achievement-level thresholds could be expressed in terms of scores on the short “test.” Such an approach might make NAEP reporting simpler, more intuitive, and more meaningful to the educators who read NAEP reports. Second, forms explicitly constructed to be used as reporting vehicles would lead to more effective item release. Market baskets would give those who use released items examples of each of the content subscales. For example, in mathematics a market basket would have some algebra items, some geometry items, and a few statistics and probability exercises. Also, release of sets of items outside the main assessment will allow analyses of those assessments to be more consistent from year to year, and therein faster and less expensive. This is consistent with the notion of a “standard” assessment.

In the case of short forms, it has been argued that such instruments might be used by schools or districts looking for a low burden, cost-effective way to evaluate their students’ performance

¹ Under some models, the short form might serve as a market basket, and vice versa.

using exercises that have solid national normative statistics. Finally, observed scores on short forms or market baskets might provide a means of releasing preliminary or initial assessment results shortly after testing, and well in advance of the final results.

For these reasons, ETS staff members welcome the opportunity to implement this important innovation. In fact, we have already conducted initial experiments with the use of market-basket metrics for reporting (Johnson, 1996); we will discuss these experiments below and will seek to build on them as we implement the market-basket study.

This chapter will be organized in the following fashion. The first section discusses market baskets and short forms, and our general approach to the study. The second section outlines some of the possible advantages and disadvantages of using a market basket in NAEP, and issues that must be resolved before market baskets can be included as routine parts of the national assessment. The third section contains an overview of our proposed design for the study surrounding market baskets. The fourth section discusses developing mathematics exercises for inclusion in the market basket, editing those items for clarity, using cognitive laboratories in the development process, pilot testing the new items, reviewing pilot-test results, and creating market-basket forms for use in conjunction with the 2000 mathematics assessment. The analysis and reporting of the results of the 2000 market-basket booklets will be discussed in Chapter 18 (Task 53).

Market Baskets and Short Forms: Our General Approach to This Study

Before moving to the main discussions in this chapter, we should explain our basic terminology and assumptions. “Market basket” has meant many things to many people. The market-basket concept is, first and foremost, a reporting mechanism. Market baskets are

sets of exercises, and assessment results are expressed in terms of a percentage of total possible scores on this set of items. The market basket might be the size of a typical test form (as is envisioned for NAEP), or might be quite a bit larger. At the most extreme, it might represent a so-called domain score. Even if it is the length of a NAEP normal test form, the market basket might or might not be administered to students.

Many people have conflated the idea of the market basket with that of the “short form.” The short form is an actual test that might be used for a variety of purposes, including linking to NAEP and as a vehicle for reporting observed scores. In one model, the market basket used for reporting would be administered to students; in such a model, *the market basket used for reporting would be a short form*. Related to this, others have suggested basing NAEP around a series of parallel test forms, rather than using the current matrix sample design.

Table 11.1 shows possible combinations of market-basket reporting and administration of short forms that might be used as part of NAEP.

Table 11.1
Reporting Possible with Different Combinations
of Market Baskets and Short Forms

	Market-Basket Reporting: Yes	Market-Basket Reporting: No
Short Form Administered: Yes	Market basket and clones used for reporting are administered to students, allowing comparison of observed and projected scores.	Like most standard achievement tests, reports involve translations of observed scores from single forms onto reporting scales.
Short Form	Scores from large-	Current situation in

Administered: No	assessment scales are expressed in terms of percentage of total score on a group of exercises.	NAEP; scores reported on an aggregate scale.
-------------------------	---	---

Given the confusion over the range of possible applications, we propose to study both reporting and administrative applications of market baskets and short forms (that is, both cells in the market-basket “yes” column). The reasons for this proposal are explained below.

The scope of work statement in the Application for Cooperative Agreement identifies two options for the market-basket study. The first approach involves developing and administering a market-basket form, while the second would be creating from released NAEP exercises a “synthetic” market basket that would be used for reporting purposes.

if we are to fully study the various range of applications of market baskets and short forms, the first approach holds more promise. Under this approach, we can assess our ability to develop forms that will serve as the basis for market-basket reports. The development effort involved will be far more constrained than those normally associated with NAEP,

. Second, including the development component in the study will allow us to build the best possible market basket from a content point of view. The items will not be limited to those previously released. Third, having a separate form administered to students in 2000 will allow for systematic comparisons of scores observed on the market basket with those imputed from links to the main NAEP assessment.

Finally, development of new exercises for release may obviate the need for wider release of exercises from the 2000 assessment at grade 4,

While market baskets and short forms are distinct, in this study *we plan to administer the market basket as a short form*. For this reason, and for the sake of ease, we use the term “market basket” to refer to both applications in the remainder of this chapter.

The Advantages and Disadvantages of Market Baskets and Issues in Need of Resolution

As mentioned above, market-basket reporting offers a number of possible advantages to the NAEP system. First, the market basket as a reporting mechanism has a compelling simplicity: Identify a set of items typical of the assessment frameworks and express assessment results in terms of a total test score on that exemplar instrument. The resulting report is simple and intuitive.² For a short enough market basket, the public is provided the complete set of items on the exemplar test, with NAEP results expressed as scores on that test.

Second, market baskets can also be a way to increase the rationality and utility of NAEP item release. Sets of questions that make up the market basket would be designed explicitly for release, thus ensuring that educators could look at sets of items that broadly represent the families of exercises in the larger main

² Although, because these instruments contain constructed-response questions scored for partial credit, in addition to multiple-choice items, the test score would not be a simple percent correct.

assessments. In addition, with sets of items pre-slated for release, analysis of the main (or standard) assessments can be identical over different cycles, leading to improvements in speed

Because of the obvious attractiveness of market basket-based reporting, ETS has prepared, as part of its NAEP redesign work (Johnson, Lazer, and O'Sullivan, 1997), experimental uses of a market-basket reporting metric. The data used for this reporting experiment were taken from the 1992 mathematics assessment. The market basket for each grade consisted of three of the released blocks of items given to that grade in 1992. It is important to note that these blocks were not designed, *a priori*, to be used as a short reporting form. The three blocks never appeared together in an assessment booklet, and were not designed to be broadly representative of the main assessment BIB. Consequently, IRT and conditioning technology were used to impute market-basket scores; no observed scores were available.

Tables 11.2 and 11.3 compare the existing reporting system with the market-basket reporting system. Table 11.2 uses the scale scores to present average proficiency, the cutpoints for the basic, proficient, and advanced achievement levels, and the percentages of students at or above each of those cut points. Table 11.3 presents the same information in terms of the test score metric on the market basket. Thus, for example, the mean score for grade 8 students would be 42 percent of the total possible market-basket score rather than a mean proficiency of 268. The advanced cut point for grade 8 would be 73 percent, rather than a scale score of 333. However, the percentage of grade 8 students above the advanced cut point is 3.8 in either metric.

Table 11.2

1992 NAEP Mathematics Assessment
Results in Terms of Scale Scores

	MEAN SCALE SCORE	ADVANCED		PROFICIENT		BASIC	
Grade		Cut Point	Percent At Or Above	Cut Point	Percent At Or Above	Cut Point	Percent At Or Above
4	220	282	3.1	249	19.1	214	61.5
8	268	333	3.8	299	23.6	262	55.6
12	300	367	2.8	336	14.9	288	63.1

Table 11.3
1992 NAEP Mathematics Assessment
Results in Terms of Market Basket
(Percent of Total Score)

	MEAN SCORE	ADVANCED		PROFICIENT		BASIC	
Grade		Cut Score	Percent At Or Above	Cut Score	Percent At Or Above	Cut Score	Percent At Or Above
4	41	80	3.1	58	19.1	34	61.5
8	42	73	3.8	55	23.6	37	55.6
12	40	75	2.8	57	14.9	33	63.1

Market-basket reporting of this type might have several important advantages. Percentages of possible scores on a defined set of items might prove intrinsically meaningful to readers: A given score would mean that students got, on average, a given proportion of a possible score on a test (for example, an “80” would indicate that students scored 40 out of 50 possible points). In addition, market baskets would likely have the effect of grounding interpretations firmly in the observed data.

While market baskets are likely to be a welcomed addition to NAEP, there are a number of possible disadvantages and issues related to their use. For example, while market baskets may be intuitive reporting mechanisms, they cannot possibly cover the range of content addressed in the main assessments. Currently NAEP assessments range from 150 to 450 minutes of assessment time; individual students are tested for only between 45 and 90 minutes (the norm is 50 minutes). Matrix sampling has been used in NAEP for an important reason: The assessments are explicitly

designed for breadth of coverage, and it is important to remember that NAEP assessments are meant to be viewed as a whole, and are not composed of redundant, parallel forms. This means that the results from the market basket might not fully reflect those based on analyses of the entire assessment. Hence, market baskets can be exemplary but not representative, and their lack of representativeness might prove confusing to readers. Content coverage is one of NAEP's virtues, and focusing on parts of NAEP that do not show such coverage may prove problematic.

There are other, more subtle, issues related to content coverage. In subjects like mathematics, one form can be constructed that contains measures of each of the five content areas covered in the assessment.³ However, each subscale will not be covered by all item types. If an extended-response question from one subscale is included in the market basket, it may necessitate not using short constructed-response questions from that subscale. Conversely, it will only be possible on a market basket to have one or two extended-answer questions. This means that subscales that in the main assessment include extended-answer questions may not have them on the market basket. Either of these scenarios will tend to weaken the relationship between the market basket and the main assessment. This might in turn cause problems in linking the market basket to the main assessment, or might result in a situation in which the observed results on the market basket tell a different story than do those of the main assessment.

There are a number of other issues that will need to be resolved before market baskets can become a regular part of the NAEP system. First, because they look like test forms, market baskets might be prone to misuse. They may not produce

³ In certain other NAEP subjects, like reading, it may not prove possible to assemble a single form of the same length as those given in NAEP that reflects the NAEP assessment and covers the domain defined in the framework. In the 50 minutes allotted for cognitive testing, a NAEP reading test taker covers at most two of the subscales. Therefore, building a single form to cover the three subscales used in grade 8 reading would necessitate an instrument longer than that used on the main assessment.

individually reliable scores, but they will be reported in a manner that will make them look like tests. This may lead people to use market baskets in inappropriate ways. Second, if we insist on linking market baskets to main NAEP scales, they may increase, rather than diminish, the complexity of NAEP. Third, while we assume that market-basket reporting will be popular with NAEP audiences, this must be confirmed through focus groups.

Finally, there is a real chance that under some uses the market baskets and the main assessment may tell contradictory stories. For example, even if the market basket has the average overall characteristics of the main assessment, results may vary for certain subpopulations. This will be a major problem if the market-basket form contains a substantial proportion of extended constructed-response questions because of the small number of these that can be included and the well-known problems of generalizability of such tasks. In general, the differences between results of market baskets and main assessments is less of an issue if one is projecting market basket scores from the main assessment, though problems may still arise.⁴ In situations where observed scores are used the problem is almost certain to become more salient.

Overview of the Study

To carry market basket-based reporting to the next level, we propose to take several steps. These steps represent the major portions of this study, and will be described in the following sections of this chapter. First, market baskets should be designed and developed to be as exemplary as possible—in both content and statistical senses—of the main assessment. Remember, earlier studies of market-basket reporting were based on groups of questions placed together *post facto* and not built to be used for

⁴ In addition, as mentioned above, such projection *adds* complexity to the system.

this purpose. The proposed design and development activity will help familiarize NAEP with the difficulties of constructing and developing market baskets, and with the types and amounts of pilot testing that prove necessary. Second, booklets consisting of the market basket will be presented to students as intact instruments.⁵ This will allow a comparison of observed scores on the market basket with those derived through the projection of scores from the main assessment. Determining the degree to which these two methods yield different answers (and they surely will differ) is a necessary precondition of using observed market-basket scores in initial reports, which could lead to substantial improvements in the rapidity in release of initial NAEP data. Third, market-basket reporting metrics should be discussed extensively with focus groups and other interested audiences. This will ensure that new NAEP reports add a valuable service to the program, and make as major and positive an impact as possible.

Design of the Study

The study will be built around the use of two market baskets in conjunction with the 2000 grade 4 assessment in mathematics. One of these forms will be composed of newly developed items and will be slated for release in a special report; the other market basket will be made up of items taken from the main NAEP assessment whose statistical qualities we know with some certitude. This latter form will be made up of secure items and not released after 2000.

Using this two-form approach will allow us to evaluate several important questions regarding market-basket use. First, building a newly developed form will allow us to gain experience in the types of pilot testing and development needed to build tests representative of the entire NAEP assessment. The development of

⁵ If IRT analysis techniques are used, the presentation of intact market baskets to students is not strictly necessary.

a form from existing NAEP items will be part of this development process: The “old market basket” will serve as a “template” that will guide and facilitate item development and review. Second, having two market baskets will allow for the comparison of results observed on different versions of the market basket. If the differences are substantial, the analysis and reporting of market-basket results may need to be adjusted and reconsidered. Related to this, using two market-basket forms will allow us to link and compare the results of both to those received from the main assessment, ensuring that projections of NAEP results onto market baskets are not idiosyncratic to a single form.

The study will have five main components or stages. First, ETS will **develop exercises** for inclusion in one “new item-based” market basket for 2000. Items will be developed during the 1998 calendar year, and pilot tested during 1999. More exercises will be developed and pilot tested than are needed in the final assessment. The assessment development process used here (and the process that will be used on other assessment development efforts) will involve the steps that have, in the past, ensured that ETS-developed instruments meet the highest standards of quality and fairness. In addition, new steps have been designed to enhance the quality of the instruments and the efficiency of the pilot testing. In the case of mathematics, a market basket constructed from existing NAEP exercises will serve as a template for development. In addition, as specified in the Application for Cooperative Agreement, **cognitive laboratories** and **special editing to ensure clarity** will be included as key components of the instrument development process. These steps should meaningfully improve the initial quality of all exercises, and thereby increase the percentage of exercises that can be used without revision after pilot testing.

After all the items are approved by NCES and the National Assessment Governing Board (NAGB), we will **pilot test all exercises, score and analyze the pilot test, review pilot-test results, and assemble one operational market basket made up of newly developed items.** The pilot testing will be conducted in February 1999. To reduce costs, this testing will be attached to the existing NAEP pilot test covered under the previously existing cooperative agreement. All revisions will be subjected to the full array of ETS reviews and external consultant reviews, and to cognitive laboratory study and special editing to ensure clarity. The operational market basket made up of new exercises will be submitted to NCES and NAGB for review and clearance in June 1999.

In the third component of the study, the market baskets will be produced as camera-ready copy, printed, and **administered in conjunction with the 2000 NAEP assessments in mathematics.** Thus, the administration of the forms will occur between January and March of 2000.

The fourth stage of the study will involve scoring and analyzing the market baskets. Analyses will be conducted to:

- look at observed overall and subgroup performance on the market basket forms,
- compare the two operational market baskets to determine the extent to which the market baskets yield similar data about the performance of students,
- link the market baskets to the main assessment, to determine whether results achieved through projecting

scores on the market baskets from main assessment data can be compared to observed scores,

- estimate the standard errors of observed scores and projected scores,
- compare the results of the market basket with those of the main assessment, to see if the findings, especially related to subpopulations, are consistent, and
- map average scale scores and achievement levels onto the market basket metric.

The final component of the study is reporting. We propose to begin work on reporting at the beginning of the grant period, by convening focus groups that will help us design reports of greatest use to NAEP audiences. Report planning will proceed through the administration and analysis periods, so that the reports can be written and released in a rapid and efficient manner. We currently plan two reports on the market-basket study: a report designed to communicate with the public about student performance on the market basket, and a research and development report outlining the results of the analytic portions of the market basket study. A complete discussion of the reports appears in Chapter 18.

The Development of Items at ETS

If the market basket is to prove a successful part of NAEP, exercises must be developed that are of the highest possible quality. As a whole, the exercises must represent, in both statistical and content senses, the overall NAEP assessment. In addition, these items will play an important public role: They must be developed to represent the content covered in grade 4 NAEP mathematics, and to communicate the general sense of what content and skills are covered in the entire NAEP assessment. ETS brings high levels of expertise and experience to the task of adding this important component to the array of NAEP instruments.

As a leader in the field of constructed-response and performance testing, ETS is singularly qualified to meet the challenges of developing a grade 4 market basket for NAEP.

As developed by ETS staff, past NAEP assessments have involved a mixture of multiple-choice and performance exercises. This allows for the measurement of a wider range of skills than is possible through use of a single item type. On the grade 4 mathematics market basket, about half of student testing time will be spent completing constructed-response exercises. The remaining time will be spent on multiple-choice exercises.

In the following sections we will discuss the development of new exercises for the grade 4 mathematics market basket. First, we will discuss ETS standards for instrument development, the ETS assessment development and review process, and the specific processes that will be used for all NAEP items. Then we will

discuss development of the mathematics exercises for the year 2000 market basket. Following this, we will describe the new cognitive laboratory and editorial processes that will become part of the reviews. Finally, we will discuss the pilot study and the plans necessary for the use of the market basket in the year 2000.

We will describe here, in some depth, the ETS assessment development and review process. This same process will be applied in other parts of this cooperative agreement—in the development of replacement exercises for U.S. history, geography, reading, and writing. Because these tasks are explicitly specified in these later sections, this text is largely repeated in Chapter 24 of the proposal.

ETS Standards for Instrument Development

Since its founding in 1947, ETS has maintained a commitment to providing services that promote precision and equity in the measurement of American education. As it has in the past, ETS will conduct NAEP in compliance with the Standards for Educational and Psychological Testing and The Code of Fair Testing Practices in Education. In addition, all ETS staff must adhere to a set of rigorous standards adopted by our Board of Trustees and set forth in the *ETS Standards for Quality and Fairness* and the *ETS Sensitivity Review Process*. These criteria are applied to all ETS-administered programs, and compliance with them has the highest priority among the ETS officers, the board, and staff.

Within the past three years, the exercise development process used for NAEP at ETS has undergone a rigorous internal and external audit to ensure that the principles described in the ETS standards are not violated. NAEP, as a project, was given extremely high marks for its adherence to these standards.

External Reviews

Content-area experts outside ETS will also review tasks and related material during this quality-assurance phase. All exercises will be reviewed, discussed, and revised at the subject-area (standing committee) meetings.

Cognitive Laboratory Study

As discussed below, all items will be subjected to think-aloud, cognitive laboratory study before use in pilot tests.

Special Editorial Review for Clarity

Based on research conducted by Jamal Abedi (Abedi, 1997) and others, ETS will institute a special editorial review specifically to ensure clarity. We propose to specially train an editor—who will

work with Dr. Abedi—to check all items for the types of unclear language or syntax that may interfere with student understanding. This review is discussed in-depth below in this chapter.

Classification of Assessment Exercises

Because an assessment is valid only if it measures the outcomes it was intended to measure, the classification of exercises is an essential step in the development process and in meeting the project's needs. Classifications affect the overall structure of the assessment, the assignment of tasks to subscales, and the reporting of scores. Only through correct classification of exercises can we ensure the close fit among assessments, frameworks, and achievement-level descriptions called for in the Application for Cooperative Agreement. The classification of exercises is challenging, given the complexity of the assessment frameworks. The challenge is increased by the fact that NAEP tasks are designed (as intended in the assessment frameworks) to be integrative and to bridge content boundaries. Therefore, ETS proposes to follow a careful, multistage process for the classification of all newly

developed exercises.

3.

4. **All classification codes will be reviewed by members of subject-area (standing) committees.** Members of these committees will be given copies of the assessment frameworks and specifications, and asked to classify all questions independently. In places where committee classifications disagree with previously assigned classifications, the matter will be discussed at the meeting of the committee. Further, committee members will be asked to perform this review every time they review a given exercise (up to four times over the course of a development cycle), to ensure that changes in classification necessitated by revisions to exercises are made. The members of the subject-area committees who served on the NAGB-sponsored planning committees that designed the assessment frameworks will play a central role in this review. These experts are in a unique position to help ensure a close fit between the intentions of the framework developers and the content of the operational assessment.

All newly developed exercises will go through the steps of this review process, to help identify areas in which the assessments need revision or augmentation to be consistent with the framework and achievement levels.

Developing Fair Assessments

ETS is committed to the development of assessments and related publications that reflect a thoughtful and balanced consideration of all peoples. *ETS Standards for Quality and Fairness* require that assessments and publications reflect the cultural backgrounds and contributions of women, minorities, people with disabilities, and other groups being assessed.

ETS policy requires that all assessments undergo a fairness review and a special empirical analysis to determine if any differential item functioning (DIF) exists. For the sensitivity review, specially trained members of the ETS staff will review individual assessment questions, the assessments as a whole, and related publications to eliminate language, symbols, or content that are considered potentially offensive. Guidelines and procedures for sensitivity review are documented in the *ETS Fairness Review Process*. For the empirical differential item functioning analysis requirement, ETS will conduct a limited DIF study as part of its analysis of the pilot test, and a more extensive study after the 2000 special study. The DIF analysis is discussed in detail in Chapters 16 and 32.

Serving Students with Special Needs

Beginning with the 1995 field test, the NAEP program embarked on a major effort to include in the assessment the maximum possible proportion of sampled students who have individualized education plans (IEPs) or are classified as limited English proficient (LEP). ETS strongly supports this objective and is committed to its

implementation to the extent that funds allow. In the 1995 field test and subsequent assessments, this effort involved the initial development and testing of a range of accommodations for these special-needs students. For students with IEPs sampled in NAEP, these accommodations took two forms. First, if a student had testing accommodations specified in his or her IEP, that accommodation was provided for the student during the NAEP field test. Second, special adaptations (for example, Braille and large-print versions of assessment books) were at times made available to students who needed them. In the case of LEP students whose native language was Spanish, Spanish-language and bilingual instruments were developed by ETS staff and used in portions of assessments. Finally, for all special-needs students, ETS developers revised the assessment scripts and administration directions in ways designed to open the instruments more fully to all.

The assessment accommodations that will be used in the market-basket study must be the subject of discussion between ETS and NCES. However, ETS proposes to follow a number of approaches in assessment development that will allow as many sampled students as possible to take part in the study. First, ETS understands that the goal of including all students in the assessment is as central to the development process as it is to the data collection process; inclusion policies cannot be effective if they are considered only *after* instruments are built. Therefore, all instruments will be developed with the needs of special-needs students in mind. For example, assessment development staff will ensure that complex language that is not essential to testing a given domain will not be included in the assessment. Furthermore, all tasks will be reviewed to ensure that they are amenable to accommodations and that as few students as possible will be excluded *a priori* because of non domain-related task attributes.

In summary, NAEP instrument development at ETS will have the full support of its expert staff, access to a wide array of ongoing diverse assessment activities, the benefit of a powerful and proven process for ensuring quality, and use of state-of-the-art technology to facilitate production. In addition, the involvement of dozens of outside teachers, experts, and interested parties from all over the country represents a major step in ensuring consensus and developing a sense of ownership for NAEP as well as acceptance of its findings.

CHAPTER 11. PREPARE FOR MATHEMATICS MARKET BASKET

Exercise Development Under This Grant

The main focus of this chapter is, of course, the development of the grade 4 mathematics market basket. However, the quality control process discussed above will be applied to all instrument development efforts subsumed under this cooperative agreement. These include the following:

- development of exercises for the mathematics market basket to be used in 2000. The market basket items will be developed in 1998 and pilot tested in 1999. This chapter discusses these activities;
- development of replacement exercises for the 2001 assessments in geography and U.S. history. These items will be developed in 1999, pilot tested in 2000, and included in the 2001 assessments. These activities are discussed in Chapters 24, 25, 26, 27, and 28;
- development of replacement exercises for the 2002 assessments in reading and writing. These items will be developed in 2000, pilot tested in 2001, and included in the 2002 assessments. These activities are discussed in Chapters 38, 39, 40, 41, and 42;
- development of materials for a special study of computer-based testing. The study will be planned in 1998, programs and exercises developed in 1999, pilot tested in 2000, and administered in 2001. These activities are described in Chapter 23; and
- AIR/ETS development of background questionnaires. This work is discussed in Chapters 8, 10, 22, and 37.

In all cases, cognitive items will be developed to match existing frameworks. All exercises will be subjected to review by a range of consultants and educators. ETS will use the processes

described above to ensure the high quality that has long marked NAEP instruments.

The ETS/NAEP Development Staff

ETS will capitalize on the measurement expertise of the members of its nationally recognized NAEP assessment development staff to ensure that the assessments reflect the frameworks, specifications, and achievement levels for each of the NAEP subject areas.

In addition, a team of senior assessment developers manages the day-to-day activities of the NAEP development group.

Each of these staff members has extensive experience in the development of forward-looking assessments,

Staff from ETS's research division who have extensive experience in the subject areas will provide consultation to the assessment development teams. These individuals' collective experience and insightful research in student learning and assessment will be key to developing the respective assessments so that they reflect the latest advances in the field.

Subject-Area (Standing) Committees

ETS has always drawn upon the expertise of external consultants including outstanding teachers, specialists in education, content-area experts, and other distinguished professionals. These consultants serve on subject-area (standing) committees. These committees are described in detail in Chapters 12 and 21.

TASK 15.

PREPARE FOR A MARKET-BASKET REPORT IN MATHEMATICS

As mentioned above, the study will proceed in five major stages, discussed in the pages that follow.

Stage 1: Develop Items For Stand-Alone Market-Basket Booklet

ETS's extensive experience in developing NAEP mathematics assessments will enable us to quickly and efficiently develop the mathematics exercises. In addition, we have a strong corporate commitment to implementing the market-basket study. Normally, we begin instrument development more than one year before pilot testing. In this case, we would like to pilot test the new mathematics exercises early in the 1999 calendar year.

Exercise development will involve the extensive reviews discussed above. The specific steps that will be followed in this study are detailed on the pages that follow.

Design for the 2000 Market Basket

The goal of this development effort is to create a grade 4 market basket. The final booklets will contain three 15-minute blocks of exercises; they will therefore be the same length as a normal NAEP mathematics assessment booklet. They will be designed to be exemplary—in both content and statistical senses—of the main NAEP assessment.

As we mentioned above, there will be two market basket forms administered in 2000. One will be composed of exercises taken from the main NAEP assessment. This form will serve as a “control,” and will not be released in the report. The second market basket, composed of newly developed exercises, will be released in full and included in the report. Because booklets in the main mathematics assessment contain three blocks of items, there will be six market basket blocks, three newly developed and three composed of exercises taken from the main assessment. Table 11.4 summarizes the blocks that will be used in 2000.

Table 11.4

Blocks Used in the Market-Basket Study in 2000

Block Number	New or Existing Items?	Comments
MB 1	New Exercises	
MB 2	New Exercises	Calculator Block
MB 3	New Exercises	Manipulatives
MB 4	Existing NAEP Exercises	
MB 5	Existing NAEP Exercises	Calculator Block
MB 6	Existing NAEP Exercises	Manipulatives

Both market baskets will be administered as stand-alone test forms. In addition, blocks from the “new item” market basket will be paired, in extra assessment booklets, with blocks from the main BIB so that statistical linkages between the main assessment and the market basket can be strengthened. The latter step is necessary if we are to project scores on the market basket from main assessment results and compare those to simple observed performance. This study will therefore necessitate adding five booklets to the 2000 mathematics spiral at grade four, two market baskets and three linking booklets. Table 11.5 shows the booklet

design that will be used in the year 2000 for the market-basket study.

Table 11.5

Booklet Design for 2000 Market-Basket Study

Booklet Number and Name	Block 1	Block 2	Block 3
Stand-Alone Market Basket 1: New Items	MB 1	MB 2	MB3
Stand-Alone Market Basket 2: Existing NAEP Items	MB 4	MB 5	MB 6
Linking Book 1	MB 1	Main NAEP Block 1	Main NAEP Block 2
Linking Book 2	Main NAEP Block 3	MB 2	Main NAEP Block 4
Linking Book 3	Main NAEP Block 5	Main NAEP Block 6	MB 3

As a working assumption, we will endeavor to make the market basket exemplary of the spread of content and exercise types and difficulties found in the main assessment. Therefore, we can write with some certainty about its characteristics. Table 11.6 gives a general description of the market basket, while Table 11.7 gives basic statistical specifications.

Table 11.6

**Tentative Design for 2000 Grade Four
Mathematics Market Baskets (Assuming 33 Total Items)**

Item Types	Multiple Choice	18
	Short Constructed Response	13-14
	Extended Constructed Response	1-2
Main Content Categories⁶	Number Sense, Properties, and Operations	14
	Measurement	6
	Geometry and Spatial Sense	5
	Data Analysis, Statistics, and Probability	4
	Algebra and Functions	4
Cognitive Categories	Conceptual Understanding	11
	Procedural Knowledge	11
	Problem Solving	11

Table 11.7

**Statistical Specifications
for 2000 Mathematics Market Baskets**

Mean Difficulty (Delta)	13.24
Standard Deviation of Deltas	2.3
Mean Discrimination (R-Polyserial)	0.629

In addition to these specifications, other conditions must be set to ensure that the market basket covers the types of exercises on the assessment as a whole. One of the three blocks in each form will require calculator use; the percentage of calculator active

⁶ Note that the numbers of items listed in the content areas are approximate. The fact that the extended-answer question will fall in one content area will, in all likelihood, reduce the number of items in that content area (because of the amount of assessment time) and lead to increases of items in other content areas.

and neutral items will mirror that found on the main assessment at grade 4. In addition, one block will include some exercises requiring the use of manipulatives (either rulers or geometric shapes). The percentage of items involving these ancillary materials will be the same as the proportion of main NAEP grade 4 mathematics items that require such materials.

As was mentioned above, there will be two market baskets in 2000, only one of which will be composed of new exercises. Thus, this proposal requires three blocks of newly developed mathematics exercises. However, unlike in past NAEP assessments, items will have to be developed to meet strict statistical assembly specifications.

Development of Pilot-Test Exercises (Task 15.1)

To ensure the development of a sufficient number of exercises to allow the assembly of a market basket that meets the content and statistical specifications described above, ETS proposes to develop and pilot test new blocks of exercises.

Normally, pilot testing the exercises needed might not be sufficient to build a market basket that meets strict statistical and content specifications. However, we hope to increase the efficiency of the development process in three ways. First, ETS mathematics assessment experts will build a market basket composed of existing NAEP items. Because we have solid statistical information on these questions, we can build a form that meets specifications. Assessment developers will use this market basket as a template that will guide and focus exercise development and should markedly increase its efficiency. Second, special editing of exercises for clarity should help ensure that items perform in pilot testing as expected. Finally, the use of cognitive laboratories should enable us

to know more about the nature and difficulty of exercises prior to pilot testing.

All items will be developed through use of the exacting processes described above. Draft items will be submitted for NCES and NAGB review by November 15, 1998.

The Mathematics Instrument Development Staff

ETS has extensive capabilities and experience in assessment development in mathematics. Among our most important experiences are the following:

After initial exercise development and review, there will be a number of special steps undertaken to ensure the quality of the mathematics exercises. These are described below.

Edit Mathematics Items for Clarity (Task 15.2)

ETS has long made efforts to ensure that cognitive exercises have been free from vocabulary or construction that interfere with accurate measurement of the desired trait. For example, it is important in a mathematics assessment that the exercises, to the extent possible, measure mathematics skills rather than reading or writing ability. While we have made substantial efforts in this area, recent research conducted by Jamal Abedi of the Center for Research on Evaluation, Standards, and Student Testing (CRESST) (Abedi, 1997) and Jay Campbell of ETS (Campbell, 1998) have suggested that further editorial work can lead to substantial improvements in the exercise development process. Such improvements should have multiple benefits, including ensuring that instruments more validly measure their intended domains, and improving the efficiency of pilot testing (as fewer items should fail for unanticipated reasons).

Abedi's work was based on the administration of "linguistically modified" NAEP mathematics forms to samples of limited English proficient (LEP) and non-LEP students. Campbell's results are based on a think-aloud protocol study involving NAEP reading exercises. Both studies had remarkably similar findings. In both cases, difficult vocabulary—even in limited amounts—confused students. It is not always easy to know which vocabulary will prove difficult. Similarly, long sentences or items, passive voice construction, conditional and prepositional clauses, and problems presented abstractly all caused problems for students that seemed not directly related to the trait being measured.

Such findings can be used to markedly improve NAEP assessments. Because of this, ETS proposes to implement a new editorial review, based on these principles, as part of all development work discussed under this application. The implementation of this review will be complex. Care must be taken to ensure that simplification of language in no way interferes with our attempts to measure educational outcomes. In other words, we do not wish to make the mathematics less challenging; we only wish to make the language more clear.

The partnership between ETS and CRESST on NAEP puts us in a unique position to design a new editorial review process. We propose to proceed in the following fashion.

_____ will head up this review effort at ETS. Her first task will be to work with _____ to design a new review process. _____ The complete description of this review process will be submitted to NCES for their approval by September 15, 1998.

Following NCES approval, _____ will train ETS editors in this new process, which will then be incorporated as a regular part of assessment development. _____ and others will

ensure that each of the new mathematics blocks receives such a review before they are either subjected to cognitive laboratory study or submitted to NCES and NAGB for clearance on November 15, 1998.

Conduct Cognitive Labs on Draft Test Booklets (Task 15.3)

ETS proposes to conduct cognitive laboratory research as part of the assessment development process in all subjects. Cognitive laboratory research will be used to discover several important attributes of questions:

- Cognitive laboratories can help determine how well the questions are understood by respondents, and how respondents' answers might be interpreted.
- Cognitive laboratories can give assessment developers greater confidence that exercises are measuring the traits they are intended to measure.
- Cognitive laboratories can help identify wording or formats that will be unnecessarily confusing or unclear to students.
- Cognitive laboratories can help identify needed refinements in scoring rubrics.
- Cognitive laboratories can help identify motivational or sensitivity problems that may not be immediately apparent to test developers.

Each newly developed mathematics item will be subjected to one-on-one cognitive laboratory work involving at least six students. Once items are revised, they will be subjected to additional cognitive laboratory research.

ETS and American Institutes for Research (AIR) staff familiar with cognitive laboratory analyses will design the specific study protocols that will be used. Both think-aloud (that is, allowing students to say what they are thinking as they are answering test

questions) and probe (protocols of questions designed to gather important information) procedures will be used. The goal of all cognitive laboratories will be to elicit practical information usable in the assessment development process.

To ensure that the cognitive labs are both efficient and useful, the first task will involve the design of working protocols.

These protocols will be submitted to NCES for review by October 1, 1998. This will enable the processes to be implemented with the new mathematics items before they are submitted for NCES and NAGB review on November 15, 1998. A memorandum describing the cognitive-laboratory results will accompany the clearance volume.

Finalize New Mathematics Market Baskets for Pilot Test (Task 15.4)

Following the extensive review process, ETS will work with members of the NAEP mathematics committee to finalize the exercises prior to their submission to NCES and NAGB for clearance. This finalization process will involve several steps:

Because of the tight development schedule, finalization activities will be completed by November 1, 1998. This will allow

for the submission by November 15, 1998 of an NCES/NAGB clearance volume containing all the exercises to be pilot tested.

**Prepare NAGB Clearance Package
for Mathematics Market-Basket Pilot Test (Task 15.5)**

Once assembly and the many reviews described above have been completed, ETS will prepare all materials needed to gain final clearance and approval of the market-basket exercises.

The clearance volume submitted for the market-basket study will include the following materials:

- a discussion of the purpose, design, and scope of the 1999 pilot test and the 2000 special study;
- a copy of each block of items to be included in the 1999 pilot test;
- answer keys for multiple-choice items and scoring guides for open-ended questions;

- content, cognitive, and achievement-level classification information on all items

- separate stimulus materials and equipment needed for students to complete each block (e.g., rulers and calculators); and

For the clearance volume for the special study in 2000, these materials will be augmented by explanations of item analysis statistics and by descriptions of criteria used to select or exclude exercises from the assessment.

The clearance package and any ancillary materials will be submitted no later than November 15, 1998. ETS will ship 11 copies of the clearance volume: five to the NAEP project officer at NCES and six to NAGB staff or board members. Because field testing should begin in January 1999, we must have clearance on field test materials no later than November 22, 1998. Printing must begin on or about that date to provide time for shipping to the field.

Prepare Camera-Ready Copy for Market-Basket Pilot Test Booklets (Task 15.6)

The preparation of camera-ready copy represents an essential step in the development of an assessment. The processes used to develop camera-ready copy have undergone a number of transformations in recent years.

In the case of the mathematics exercises, production of camera-ready copy will begin after the market baskets have been assembled and have been given their initial committee reviews. This production will involve layout of text and production of

artwork. In addition, production work will begin on any ancillary materials that students will use in the assessment but that will not be printed in assessment booklets.

As mentioned earlier, there are several reviews of camera-ready copy designed to ensure quality

These reviews are necessary to ensure that NAEP assessments are free from error.

Stage 2: Pilot Test Short-Test Booklet, Analyze Pilot-Test Results, and Select Exercises for Special Study

Pilot Test in 1999 (Task 15.7)

After clearance, books will be printed and shipped to the field. The pilot test of the two mathematics market baskets will be held in January and February 1999. Each book will be given to 500 students, so we may generate reliable item-analysis data. For the sake of efficiency, these books will be spiraled in with other mathematics field test booklets that have been developed under the existing NAEP cooperative agreement.

Return Pilot-Test Booklets to NCS and Score Pilot Test (Task 15.8)

After the pilot test, market-basket booklets will be returned to NCS for scoring. There multiple-choice questions will be scanned and constructed-response questions will be scored on the image system. The scoring process is described in depth in Chapter 14. The

description will not be repeated here, as the same procedures will be followed for the mathematics pilot test. Pilot-test scoring will take roughly three days, and will involve about 20 scorers. Scoring will take place in March 1999. Following scoring, NCS will prepare data tapes with assessment results, and send these to ETS for item analysis.

Analyze Pilot-Test Results (Task 15.9)

Upon receipt of the data from NCS, ETS analysts will conduct pilot-test item analysis. This analysis will result in the following data.

Dichotomous Items. Multiple-choice and dichotomously scored constructed-response items will be analyzed using standard procedures that result in a report for each item,

Polytomously Scored Items. Enhanced procedures, developed by ETS staff for previous NAEP assessments, will be employed for polytomously scored items (those having more than two outcome categories). First, methods parallel to those for dichotomously scored items will result in values reported for each distinct response category for the item. For example, a constructed-

response item with four response categories will result in six categories (not reached and omitted in addition to the four response categories). Another set of statistics will result from mapping the response categories (excluding not-reached) into a new set of categories reflecting the scoring rubric for the item. A constructed-response item with ordered categories, for example, will be mapped into the integers in a corresponding order.

For these items, the following statistics, analogous to those for dichotomously scored items, will be computed:

Differential Item Functioning Analyses. These analyses will be carried out on each block of items that is pilot tested, using the standard procedures described in detail in Chapter 16. The only demographic variable for which the sample sizes will be large enough to do these analyses, however, will be gender. Hence, these analyses will be carried out only with female students as the focus group and male students as the reference group.

Select and Revise Exercises for 2000 Special Study of Market Basket (Task 15.10)

Based on the review of pilot test results, exercises will be selected to for use in the “new” market-basket form in 2000. In deciding which items move forward from the pilot test to the special study,

ETS development staff use a variety of criteria; these criteria will differ, in minor ways, from those commonly used in NAEP assessments. In some cases these criteria are also used to evaluate the need for revision of questions that are to be included, and to determine what the revisions should be. However, because we are attempting to build a market basket to meet strict statistical specifications, we will endeavor to hold post-pilot-test revisions to an absolute minimum. The interrelationships between these criteria are complex, and decisions about the inclusion or exclusion of items and tasks are made by considering a large number of factors in concert. The most important of these criteria are described below.

- **First and foremost, the pool of items selected for the operational assessment must meet content and statistical specifications created for the market basket.** These specifications, described above, will serve as the most important criteria ETS development staff will use in choosing questions. Within this general criterion, all attempts are made to choose exercises that meet the other selection criteria.
- **The items must cover the range of NAGB achievement-level descriptions.** ETS test development specialists and subject-area committees will review all items to ensure that they measure, to the maximum extent possible, knowledge and skills in a manner consistent with NAGB achievement-level descriptions.
- **Information from the pilot-test scoring is used in decisions to include or exclude constructed-response tasks, and as a guide to needed revisions.** Student responses are one important guide to how well an item is functioning. Tasks that do not yield useful responses may be excluded from the assessment, or may be revised in such a way as to lead students to better responses. When available, rater agreement data may also be used to inform decisions about retaining or excluding a task.

- **All exercises are reviewed against statistical indices prior to their inclusion in an operational assessment.**

Various statistical indices are used to determine whether an item should be used in an operational assessment.

Statistical factors are normally but one of the criteria assessed when crafting an operational assessment; in most cases, the priority in building the assessment is conformity with frameworks and specifications. However, if the market basket is to contain items exemplary of the range of difficulty present in the assessment, items must be chosen that fit the statistical specification described under Stage 1 of this plan. The statistical criteria that will be used by ETS/NAEP development staff in developing the market basket are described below.

- **Item discrimination.** The degree to which an exercise discriminates between students of different abilities (as measured by the r-biserial and r-polyserial statistics) is an important indicator of exercise performance. On the whole, ETS staff endeavor to ensure that questions selected have, wherever possible, r-biserials or r-polyserials above . In practical terms, items with higher r-biserials or r-polyserials are always preferred to those with lower values, other criteria being equal. In any case where it is necessary (for compliance with assessment specifications) to use an item with an r-biserial or r-polyserial lower than , development staff engage in a thorough review of the item. Using the item analysis as a guide, development specialists attempt to determine if minor revisions to the items are likely to yield better statistical results. In the case of a multiple-choice item, this often involves examining the one or two of the incorrect response options that seem to be attracting able students. For constructed-response questions, low r-polyserials can be caused by problems at specific points in scoring guides. These problems can often be fixed by trained assessment specialists.
- **Difficulty.** Unlike other NAEP assessments, for the market basket assessment developers will work from a statistical specification that indicates the average item difficulty and

the standard deviation of difficulties that will serve as targets for development. Items will be selected that meet these criteria.

- **Not-reached rates.** In the case of the mathematics market-basket assessment, past data will give us good information on speededness, so the market-basket lengths described above should be appropriate. However, if any item or task has a "not-reached" (defined as students who do not answer a given question and no other questions that appear later in a block) percentage higher than $\frac{1}{3}$, ETS development staff are instructed to treat interpretation of the item statistics with extreme caution. Wherever possible, assessment developers are instructed use questions with "not-reached" rates below $\frac{1}{3}$ percent rather than those with higher rates, all other criteria being equal. This is because we have higher confidence in the item analysis statistics of exercises that all students have had a chance to complete.
- **Omission rates.** The omission rate may be defined as the percentage of students not answering a given item or task who do answer a question that appears later in the assessment. While it is preferable that all omission rates be low, widespread use of constructed-response questions leads to the necessary use of some tasks with higher than optimal omission rates. ETS development staff use the omission rate in two ways. First, wherever possible tasks with omission rates below $\frac{1}{3}$ percent are used rather than those with higher rates. Second, wherever it is necessary to use tasks with higher omission rates, ETS development staff use all available information gathered from the field test to implement item revisions designed to maximize student response rates.
- **Scoring reliability.** For constructed-response questions, evidence from the pilot test of scoring reliability weighs heavily in the exercise selection process. Exercises that can be scored reliably will be selected before those that cannot. Any exercise with low scorer agreement will be used only if assessment developers are confident that revisions to scoring guides can lead to better performance.

- **Performance of options on multiple-choice items.**

Assessment developers will look for multiple-choice exercises in which students use all answer options. Options that attract no students indicate suboptimal item functioning, and lead to the exclusion of the item from the operational assessment or the revision of the unused option.

- **Spread of scores on constructed-response questions.**

Developers will prefer constructed-response questions on which student responses are spread over the score scale to those in which they all cluster at one or two score points. However, this criterion is less important than content coverage. Certain exercises will be used to measure key content, even if they display a suboptimal spread of difficulties.

ETS assessment specialists will review all pilot-test exercises against the criteria described above. In selecting items, preference will be given to exercises that can move from pilot test to special study without revision. If such exercises can fill all content holes in the assessment, no exercises will be revised. Of course, content concerns will be primary, and it is likely, in trying to meet the specifications for the assessment, that some degree of exercise revision will be necessary. However, the new pilot-test review steps (that is, editing for clarity and cognitive laboratory study) should reduce the proportion of exercises that need review.

Review All Exercises and Blocks Selected for Operational Use (Task 15.11)

All exercises selected for use in special market-basket study—whether used in their pilot-test form or revised—will undergo extensive reviews *after* the pilot test as well as before. The review steps have been described extensively earlier in this chapter. The in-depth descriptions of those reviews will not be repeated here. Suffice it to say that the market-basket will be reviewed for content appropriateness , for

editorial accuracy, for clarity, for fairness, and, at the camera-ready stage, for accuracy. In addition, the operational selections will all be reviewed and approved by members of the NAEP mathematics subject-area committee, as well as by other consultants who work on the project. Any exercise that is revised will be subjected to cognitive laboratory study before inclusion in a NAGB clearance package.

These activities will be completed for the market-basket blocks by June 30, 1999.

Prepare NAGB Clearance Package of Market-Basket Special Study Blocks for Use in 2000 (Task 15.12)

Using procedures similar to those for the pilot test, ETS will prepare all materials needed for the formal clearance and approval process. For the sake of efficiency, these materials will be combined with the 2000 mathematics clearance volume being prepared under the existing cooperative agreement. Therefore, no special activities will be described here.

The clearance package and any ancillary materials will be submitted no later than July 31, 1999. Because operational testing should begin in January 2000, we must have clearance on field test materials no later than September 1, 1999. Printing must begin on or about that date to provide time for shipping to the field.

Throughout the process, ETS staff will be available, if necessary, to meet with NCES and NAGB staffs for reviews of materials. At any time in the clearance process, ETS is prepared to clarify any ambiguities and answer questions to expedite the process.

Stage 3: Administer Special Study

Prepare Camera-Ready Copy of Market-Basket Booklets for Use in 2000 (Task 15.13)

The new process for producing camera-ready copy of assessment booklets has been described in depth earlier in this chapter. We will not repeat the details of the process here; the essential steps followed will be the same as those used during the pilot test.

Production of camera-ready copy for the mathematics market-basket will begin after the blocks have been assembled and been given their initial committee reviews. The production process will make use of the photocomposition done for the pilot test. This creation of camera-ready copy involves layout of text and production of artwork. There are several reviews of camera-ready copy that are designed to ensure quality.

These reviews are necessary to ensure that NAEP assessments are free from error.

The first version of camera-ready copy will be ready by June 30, 1999, and to be included in the NAGB clearance volume. After clearance, camera-ready copy will be revised in line with NAGB and NCES suggestions. Final camera-ready copy will be sent from ETS to NCS for printing no later than September 2, 1999.

Administer Study (Task 15.14)

As mentioned above, the market-basket will be administered to students between January and March 2000. Two thousand students will take each of two market-basket forms (one “new” form and one “template” form). In addition, 2,000 students will complete booklets which will allow the linking of the market basket to the main assessment. For details of the design, see Table 11.3 and Table 11.4 above.

Stage 4: Score and Analyze Market Basket

Scoring (Task 15.15)

The market basket will be scored in conjunction with the 2000 NAEP assessments in mathematics and science. Procedures for that scoring are described in Chapter 14. In addition, scoring activities for the market basket have been budgeted under Task 28.

Analysis (Task 15.16)

Analysis of the market basket is described Chapters 16 and 18, which is where funds for the tasks have been included. In general, analyses will look at student performance on the market basket, the reliability of that instrument, and the links between the market basket and the main assessment.

Analysis plans will become clearer as the study is implemented. We will deliver to NCES an analysis plan by October 1, 1999.

Reporting (Task 15.17)

We plan two market-basket reports. The first will be a report of market-basket results aimed at the general public. The second will be a research and development report that will summarize what

has been learned from the study. Both reports are described in Chapter 18 (see Task 52).

Finding ways to make NAEP results more meaningful to a broad public is a key goal of NAGB redesign. Market baskets may be one way of accomplishing such improvements. ETS looks forward to taking our first steps into this important new arena.

CHAPTER 12.

NAEP MATHEMATICS, SCIENCE, READING, AND WRITING STANDING COMMITTEES

Tasks 16 and 17

EXECUTIVE SUMMARY

Objective: To ensure that subject-matter specialists oversee the assessment process. To meet this objective, Educational Testing Service (ETS) will:

- work with standing subject-area committees comprised of teachers, teacher educators, researchers, and policymakers to ensure that the assessments appropriately and accurately measure what students know and can do,
- work with these standing committees to ascertain the types of reports needed to make NAEP data accessible and useful to different audiences,
- appoint a facilitator for each standing committee who will review the agenda and help lead group discussions;
- meet periodically with each standing committee throughout the assessment cycle so that advice can be given in a timely manner, and
- prepare briefing books prior to each standing committee meeting and ensure that meeting minutes are disseminated within three weeks following each meeting.

Overview of Tasks 16 and 17

The National Assessment of Educational Progress is the nation's primary tool for measuring students' educational achievement. It is, therefore, of paramount importance that each assessment reflect policies and practices that are educationally sound. To ensure that this goal is met, Educational Testing Service (ETS) has made use of educators from different walks of life. Teachers, teacher educators, researchers, and policymakers have brought a wide range of experiences to NAEP. They have helped ensure that the assessments appropriately and accurately measure what students know and can do, as well as have provided advice on the types of reports needed to make NAEP data accessible and useful to different audiences.

In recent years, ETS has sought advice and expertise from two committees during each NAEP assessment cycle: an instrument development committee and a standing committee. While these two committees overlap in membership, the former is designed to advise on instrument development and the latter to provide content, technical, and procedural advice. These roles have become blurred, with individuals who are members of the standing committee taking on some of the roles of members of the instrument development committee. ETS recognizes the need to ensure consistency across the NAEP cycle and to tap into the collective experiences of the education community. We believe that each subject assessment would profit if one group maintained a consistent oversight function throughout all phases of the project (including instrument development, analysis, and reporting). We further feel that one enhanced standing committee can efficiently play the roles currently performed by the two bodies. Therefore, given the current blurring of roles and amount of overlap between the two committees, ETS would like to suggest combining them to form a cohesive group that will provide continuity across the NAEP

assessment cycle, be connected intimately to the instrument, and be able to provide technical and procedural advice.

If agreed to by the National Center for Education Statistics (NCES), four committees are proposed:

- NAEP Mathematics Committee
- NAEP Science Committee
- NAEP Reading Committee
- NAEP Writing Committee

These committees will be made up of leading educators in their fields. Some members will be recruited from the groups that developed the NAEP subject-area framework governing an assessment. NAEP subject committee meetings will be held in Washington, DC. One member from each committee will be invited to serve as a facilitator. The selection of the facilitators will be based on their broad-based involvement with every phase of the project and their availability to participate fully in the process. The agenda for each meeting of these committees will be determined by NCES and ETS. ETS, working with the facilitators, will convene the committees and provide technical and administrative support to them as needed. The facilitators will review the agenda, help lead group discussions, and serve as the initial points of contact with other committee members when immediate feedback is needed between meetings. Briefing books will be prepared prior to each meeting and minutes will be sent within three weeks following each meeting. ETS will also communicate with the facilitators of the committees whenever issues arise that require immediate feedback.

TASK 16.

NAEP MATHEMATICS COMMITTEE AND NAEP SCIENCE COMMITTEE

Instrument development committees are already in place for the 1999 science and mathematics field tests. They consist of members who have served on one or more phases of the NAEP assessment cycle as well as new members. A combined mathematics/science standing committee of six people from the mathematics community and five from the science community is also in place. ETS proposes, with NCES's approval, to merge these two committees such that the members of the Mathematics Instrument Development Committee and the mathematics representatives on the Mathematics/Science Standing Committee will become members of one committee—the NAEP Mathematics Committee. Similarly, members of the Science Instrument Development Committee and the science representatives on the mathematics/science standing committee will also become members of one committee—the NAEP Science Committee. Thus, a NAEP Mathematics Committee and a NAEP Science Committee will be formed, each with a proposed membership of 15. Committee members will be nominated by ETS and approved by NCES. ETS would like to suggest that the current members be retained, since collectively they have been involved with framework development, instrument development, report planning, writing, and review, and have provided content, technical, and procedural advice.

Because mathematics and science are on the same administration schedule, the NAEP Mathematics Committee and the NAEP Science Committee will each meet twice before the 2000 assessment to review the scoring guides and student responses from the 1999 field test and to review the final forms prior to government clearance. They will also advise ETS on matters that pertain to the overall success of each program. These meetings are

funded in the current budget for the existing cooperative agreement covering NAEP activities between 1996 and 1999.

The NAEP Mathematics Committee and NAEP Science Committee will

consider the overall direction of the project. The members of each committee will also review plans for reports arising from the 2000 assessments and will review data and give advice on report content and dissemination.

ETS feels strongly that the presence of one committee per subject area will streamline oversight of the entire NAEP cycle and give the members of the committee clearly defined purposes and goals concerning their role in the project.

Proposed NAEP Mathematics Committee Members

The NAEP Mathematics Committee will be made up of the 13 members of the Mathematics Instrument Development Committee and five members of the Mathematics/Science Standing Committee (four of the standing committee members are also on the instrument development committee). One new member has been proposed. The names and affiliations of the proposed members are shown below together with information relating to their past NAEP connections.

Leslie Djang, Sandy Run Middle School, PA

*NAEP Mathematics Instrument Development Committee
(1998 -Present)*

John Dossey, Illinois State University, IL

*NAEP Mathematics Instrument Development Committee
(1992-1996, 1998 -Present)*

NAEP Mathematics Standing Committee (1995-1996)

NAEP Mathematics/Science Standing Committee (1996-Present)

Lucy Garner, Los Angeles Center for Enriched Studies, CA

CHAPTER 12. MATHEMATICS, SCIENCE, READING, AND WRITING COMMITTEES

*NAEP Mathematics Instrument Development Committee
(1998 -Present)*

Bill Hopkins, Texas Education Agency, TX
*NAEP Mathematics Instrument Development Committee
(1998 -Present)*

Audrey Jackson, Parkway School District, MO

NAEP Mathematics Instrument Development Committee
(1992-1996, 1998 -Present)

NAEP Mathematics Standing Committee (1995-1996)
NAEP Mathematics/Science Standing Committee (1996-Present)

Jeane Joyner, North Carolina Department of Public Instruction,
NC

NAEP Mathematics Instrument Development Committee
(1998 -Present)

Constance Kelly, Bloomfield Middle School, MI

NAEP Mathematics Instrument Development Committee
(1998 -Present)

Mary Lindquist, Columbus State University, GA

NAEP Mathematics Instrument Development Committee
(1992-1996, 1998 -Present)

NAEP Mathematics Standing Committee (1995-1996)
NAEP Mathematics/Science Standing Committee (1996-Present)

Rochelle Newman, West Windsor-Plainsboro Middle School,
NJ

NAEP Mathematics Instrument Development Committee
(1998 -Present)

Ismael Olivas, Socorro High School, TX

NAEP Mathematics Instrument Development Committee
(1998 -Present)

Christopher Olsen, George Washington High School, IA

NAEP Mathematics/Science Standing Committee (1996-Present)

Deborah Romanek, Nebraska Department of Education, NE
(new)

Catherine Phillips, Tollgate Grammar School, NJ

NAEP Mathematics Instrument Development Committee
(1998 -Present)

Charles Watson, Arkansas Department of Education, AR

NAEP Mathematics Instrument Development Committee
(1998 -Present)

Zalman Usiskin, University of Chicago, IL

NAEP Mathematics Instrument Development Committee
(1992-1996, 1998 -Present)

NAEP Mathematics Standing Committee (1995-1996)

NAEP Mathematics/Science Standing Committee (1996-Present)

Proposed NAEP Science Committee Members

The NAEP Science Committee will be made up of the 12 members of the Science Instrument Development Committee and five members of the Mathematics/Science Standing Committee (three of the standing committee members are currently members of the instrument development committee). One other proposed member has been listed. The names and affiliations of the proposed members are shown below together with information relating to their past NAEP connections.

Michael Burton, Discovery Middle School, ND

NAEP Science Instrument Development Committee (1998 - Present)

Lucy Caballero, Hereford Junior High School, TX

NAEP Science Instrument Development Committee (1998 - Present)

Audrey Champagne, State University of New York, NY

NAEP Science Framework Committee (1991 -1992)

NAEP Science Instrument Development Committee (1992-1996, 1998 -Present)

NAEP Science Standing Committee (1995-1996)

NAEP Mathematics/Science Standing Committee (1996-Present)

Russ Conner, Cranbrook Kingswood School, MI

NAEP Science Instrument Development Committee (1998-Present)

Robert Corell, Princeton High School, NJ

(new)

Patricia Dung, LA Educational Partnership, CA

NAEP Science Instrument Development Committee (1992-1996, 1998 -Present)

NAEP Science Standing Committee (1995-1996)

NAEP Mathematics/Science Standing Committee (1996-Present)

Ed Hendry, New Hampshire Department of Education, NH

NAEP Science Instrument Development Committee (1998-Present)

Michael Jojola, Isleta Elementary School, UT

NAEP Science Instrument Development Committee (1995-1996, 1998 -Present)

Brett Moulding, Utah State Department of Education, UT

NAEP Science Instrument Development Committee (1998-Present)

Kelly Poling, Logan-Hocking Local School District, OH

NAEP Science Instrument Development Committee (1998-Present)

Senta Raizen, National Center for Improving Science
Education, DC

NAEP Science Framework Committee (1991 -1992)

*NAEP Science Instrument Development Committee (1992-
1996, 1998 -Present)*

NAEP Science Standing Committee (1995-1996)

NAEP Mathematics/Science Standing Committee (1996-Present)

Realista Rodrigues, Pimmit Hills School, VA

*NAEP Science Instrument Development Committee (1992-
1996)*

NAEP Science Standing Committee (1995-1996)

NAEP Mathematics/Science Standing Committee (1996-Present)

Elise Russo, New York State Education Department, NY

*NAEP Science Instrument Development Committee (1998-
Present)*

Gerald Weaver, University City Cluster Office, PA

*NAEP Science Instrument Development Committee (1992-
1996, 1998 -Present)*

NAEP Science Standing Committee (1995-1996)

Gerald Wheeler, National Science Teachers Association, VA

NAEP Mathematics/Science Standing Committee (1996-Present)

TASK 17.**NAEP READING COMMITTEE AND
NAEP WRITING COMMITTEE**

Instrument development committees are already in place for the 1999 reading field test and the 1998 writing main assessment. They consist of members who have served on one or more phases of the NAEP assessment cycle. A reading/writing standing committee, consisting of five people from the reading community and five people from the writing community, is also in place. ETS proposes, with NCES's approval, to restructure these two committees so that members of the Reading Instrument Development Committee and the reading representatives on the Reading/Writing Standing Committee become one committee—the NAEP Reading Committee. A NAEP Writing Committee will be constructed in a similar way. Thus, a NAEP Reading Committee and a NAEP Writing Committee will each be formed with a proposed membership of fifteen. The committee members will be nominated by ETS and approved by NCES. ETS would like to suggest that the current members be retained since collectively they have been involved with framework development, instrument development, report planning, writing, and review, and have provided content, technical, and procedural advice.

The NAEP Reading Committee will meet before the 2000 assessment to review the scoring guides and student responses from the 1999 field test, review the final forms prior to clearance, and provide guidance on 1998 report writing. They will also advise ETS on matters that pertain to the overall success of the program.

The NAEP Reading Committee will then meet _____ a year for the next three years of the cycle and _____ in the final year. In addition to providing guidance on assessment development, members of the committee will be concerned with the reports that arise from the 2000 and 2002 assessments.

The NAEP Writing Committee will meet _____ in the course of 1998 and 1999 to provide guidance on report writing following the 1998 assessment.

_____ The NAEP Writing Committee will then meet _____ for the next three years of the cycle and _____ in the final year (2000 - 2003). During this time period committee members will help prepare questions to be pilot tested in 2001, review scoring guides and student responses from the 2001 NAEP pilot test, review the final forms prior to clearance, and provide guidance on report writing following the 2002 NAEP assessment. Committee members will also advise NCES and ETS on matters that pertain to the overall success of the program.

Proposed NAEP Reading Committee Members

The proposed NAEP Reading Committee comprises one former and ten current members from the Reading Instrument Development Committee and five members from the Reading/Writing Standing Committee (four also serve on the instrument development committee). Three new members have also been proposed.

Peter Afflerbach, University of Maryland, MD
NAEP Reading Instrument Development Committee
(1996-Present)

Rosalinda Barrera, New Mexico State University, NM
NAEP Reading Instrument Development Committee
(1996-Present)

Susan Biggam, Department of Education, VT
(new)

Janice Dole, University of Utah, UT
NAEP Reading Instrument Development Committee
(1996 -Present)

John Guthrie, National Reading Research Center, MD
NAEP Reading Instrument Development Committee
(1996-Present)
NAEP Reading/Writing Standing Committee (1997-present)

Violet Harris, University of Illinois at Urbana, IL
NAEP Reading Instrument Development Committee (1996-Present)
NAEP Reading/Writing Standing Committee (1997-present)

Janet Jones, Charles County Public Schools, MD
NAEP Reading Instrument Development Committee (1992-Present)

Judith Langer, State University of New York, NY
NAEP Reading Instrument Development Committee (1992-1994)
NAEP Reading/Writing Standing Committee (1997-Present)

Patricia McGonegal, Mount Mansfield Union High School, VT
NAEP Reading Instrument Development Committee (1996-Present)

Jane C. Miller, Barnes Elementary School, VT
(new)

Susan Neuman, Temple University, PA
NAEP Reading Instrument Development Committee (1992-Present)
NAEP Reading/Writing Standing Committee (1997-Present)

Jack Pikulski, University of Delaware, DE
NAEP Reading Instrument Development Committee (1992-Present)

Timothy Shanahan, University of Illinois at Chicago, IL
NAEP Reading Instrument Development Committee (1998-Present)
NAEP Reading/Writing Standing Committee (1997-Present)

Peter Winograd, University of New Mexico at Albuquerque, NM
(new)

Karen Wixson, University of Michigan, MI
NAEP Reading Instrument Development Committee (1996-1998)

Proposed NAEP Writing Committee Members

The proposed NAEP Writing Committee will be made up of nine members from the Writing Instrument Development Committee and four members from the Reading/Writing Standing Committee (three of whom also serve on the instrument development committee). Five new members have also been proposed.

Arthur Applebee, SUNY-Albany, NY

NAEP Writing Framework Committee (1989 -1990)

NAEP Writing Instrument Development Committee (1990-1992, 1996 -Present)

NAEP Writing Planning Committee (1996-1997)

NAEP Reading/Writing Standing Committee (1997-Present)

Katherine Au, University of Hawaii, HI

NAEP Writing Instrument Development Committee (1996 - Present)

Alyse Eidman-Aadahl, National Writing Project, CA

NAEP Writing Instrument Development Committee (1996 - Present)

NAEP Writing Planning Committee (1996-1997)

NAEP Reading/Writing Standing Committee (1997-Present)

Sylvia Flores, Artesia Public Schools, NM

NAEP Writing Instrument Development Committee (1996 - Present)

NAEP Writing Planning Committee (1996-1997)

Kris Gutierrez, UC-Los Angeles, CA

NAEP Writing Instrument Development Committee (1990-1992, 1996 -Present)

NAEP Reading/Writing Standing Committee (1997-Present)

Geraldine Guttwein, Harrisburg Area Community College, PA

NAEP Writing Instrument Development Committee (1996 - Present)

Charlotte Higuchi, CRESST, CA

NAEP Writing Instrument Development Committee (1996 - Present)

NAEP Writing Planning Committee (1996-1997)

Brian Huot, University of Louisville, KY
NAEP Writing Instrument Development Committee (1996 - Present)

Miles Myers, National Council of Teachers of English, CA
NAEP Writing Instrument Development Committee (1990-1992)
NAEP Reading/Writing Standing Committee (1997-Present)

Patricia Porter, Texas Education Agency, TX
NAEP Writing Instrument Development Committee (1996 - Present)

Arlene Sinding, Hillsborough High School, NJ
(new)

Edward Uehling, Valparaiso University, IN
(new)

Peter Valenti, Fayetteville State University, NC
(new)

Daisy Vickers, North Carolina Department of Public Instruction, NC
(new)

Darian Walker, Pinellas County Schools, FL
(new)